ON EVOLUTIONARY THEORY, INFERENCE, AND SIMULATION: A GENEALOGICAL PERSPECTIVE

by

Scotland Charles Leman

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Michael Lavine, Supervisor

_____
Dr. Yuguo Chen

_____
Dr. Sayan Mukherjee

_____
Dr. Mark Huber

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2007

ABSTRACT

(Statistics)

ON EVOLUTIONARY THEORY, INFERENCE, AND SIMULATION: A GENEALOGICAL PERSPECTIVE

by

Scotland Charles Leman

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Michael Lavine, Supervisor

_____
Dr. Yuguo Chen

_____
Dr. Sayan Mukherjee

_____
Dr. Mark Huber

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2007

Abstract

This thesis discusses evolutionary inference from both a modeling perspective and the algorithms associated with performing statistical inference.

Genetic data (DNA) takes on a nontraditional form in that a single observation encompass at least hundreds of base pairs and is nonnumeric in nature. Beyond this fact, DNA from individuals that share a common ancestry have similarities in their genetic makeup, so the notion of independent and identically distributed samples does not hold. In turn, a complex network of associations must be employed when modeling the data.

The complexities involved in the modeling procedure directly relate to the complexities involved when reconstructing likelihood functions, or posterior distribution. Many computational methods used during statistical inference involve the idea of drawing samples from proposal distributions. However, such proposal distributions are difficult to construct so that their probability distribution match that of the true target distribution, in turn hampering the efficiency of the overall sampling scheme.

We will describe a general approach to modeling the evolutionary past. Within this framework, we will discuss specific models which address particular phenomena (speciation, introgression and paracentric inversions) which relate to genomic data. The latter part of this thesis will address two simulation methods used for statistical inference. The first will pertain to direct likelihood construction under an Importance Sampling (IS) framework and the second will address a Markov Chain Monte Carlo (MCMC) procedure for posterior sampling.

Acknowledgements

I would like to express my deep gratitude to those have been influential through out my tenure as a graduate student. First of all, I would like to thank my advisor Michael Lavine. While these topics can never be decoupled, Michael has guided me not only as a scientific researcher but also as a statistical philosopher.

I would like to thank Tom McGrath for driving me out to Durham. That drive constantly reminds me that even when I run out of gas, I will always make my destination. To my most reckless friends, Joey Teran and Pat McGrath, thank you for the constant entertainment and making me always worry about what might happen next.

My Brother, Steve Leman, I would like to thank above all for always setting the bar high. It was his footsteps that have led me to pursue academics. It was his high standards that made me give my best effort. While we have always competed, he has set the pace and has provided me with the determination and focus to always meet my goals.

Leanna House, my best friend and companion. I thank you for always standing by me and for your constant voice of encouragement. Your sunny disposition and love for life has made my days brighter and for that I'll always love you.

To my surrogate family, Chris and Sara Ecclesine, I wish to thank you for all of your wonderful generosity.

My friends at ISDS have been plentiful and I would like to thank you all for the wonderful moments that we have shared. Special thanks is reserved for Chong Tu, Natesh Pillar and Eric Vance. Chong, thanks for setting such a good example for me in my first year. Natesh, thank you for always supplying statistical food

for thought. Eric, thanks for all the comic relief.

Without all of you, this process would have been extremely difficult, if not impossible. Thank you all for making this such a wonderful experience for me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The field of statistics has grown tremendously in the last 70 years, dating back to the days of Fisher and being motivated primarily by genetic and biological problems. However, the field of genetic inference on a molecular level is just getting past its infancy. The principal reason for this is that biologists have recently begun to refine methods of data acquisition, and molecular data are now plentiful. Unfortunately, these data are not of the traditional shape and form that the field of multivariate statistics has become accustomed to dealing with. Primary problems stem from the fact that nucleotide sequence data is nonnumeric, massive in scale (the human genome is made up of approximately 6 billion bases), and the notion of independent and identically distributed samples often does not apply. While the first of these three problems is easily tackled by numeric summaries of the sequence data, the latter two are problematic and are confounded by each other. That is, models based on genomic associations are complex and the scale of the data creates computational demands that are only surmountable through large networks of computers or supercomputers. This thesis is a cumulative work in understanding particular numeric summaries from

genomic data, the patterns which arise in DNA sequences and how they related to ancestral relationships between genes, models for genealogical inference and the computation which arises in handling such inference problems.

1.1 Thesis outline

This thesis will comprise modeling, methods and analyses on real data.
Chapter 3 will introduce the models which will be studied. This work is an extension of the original coalescent work summarized in Kingman (2000) and has been developed by Marcy K. Uyenoyama, and myself. The bulk of chapter 3 is based on a probability generating function which computes the probability of mutation counts from DNA data which can be found in appendix A. This work was previously introduced by Uyenoyama and Takebayashi (2004). Appendix A begins with laying out the theoretical constructs for modeling, while chapter 3 sections 3.2.1, 3.2.7 and 3.2.12 will specifically show how the modeling framework can be applied to precise structured models. Chapter 3 will also discuss computational issues involved with these models and motivates the sampling methodology discussed in subsequent chapters.

In chapter 4 we will describe an Importance Sampling (IS) technique which is described in the paper by Leman *et al.* (2005). This method greatly reduces the computational demands required by the exact computation of probability generating functions as described in chapter 3. Chapter 4 will illustrate the proposal distributions used for the IS method (see section 4.2) and will describe the tuning procedures required for the method (see sections 4.3 and 4.3.3).

Chapter 5 provides a case study for analyzing real data under the model described in section 3.2.7. This case study will employ the importance sampling method described in chapter 4. Comparisons to exact methods as described in chapter 3 are also detailed.

Chapter 6 will describe the Evolutionary Forest (EF) algorithm. The EF algorithm is a novel Markov chain Monte Carlo procedure for constructing the posterior distribution of the interesting parameters in the models described in chapter 3. Instead of examining the posterior distribution for a single genealogy and parameters, the EF method replaces the space in which the genealogy lives with a forest of genealogies. In chapter 6 we will describe both the distribution on forest space and the algorithm used for sampling. Theoretical results will also be shown. Chapter 6 will conclude with a case study and a comparison to alternative methods.

Chapter 2

An Historical Perspective

We begin with a survey of the field of population biology which dates back to 1930 when Fisher laid out the foundations of the field and explored concepts of genetic drift and how it governs evolution. Through this historical perspective, we will motivate coalescence theory, which serves as a powerful device under which real data can be analyzed.

2.1 Genetic drift and a forward perspective

Genes are passed down from generation to generation through breeding. Each time a new individual is introduced into the population, small differences between the individual and its parent's genome exist. These difference can occur due to selection, the process in which particular genes are selected based on a propensity to increase fitness, or neutral drift. This process of neutral drift is a process in which random mutations are introduced and aren't influenced by outside selective pressures. We will often refer to neutral drift as genetic drift or simply as drift.

Neutral drive is governed my a constant neutral rate of mutation, which we will denote as u. It is the neutral rate of mutation that is responsible for how fast drift occurs. Neutral drift is responsible for a vast quantity of the species we see today. That is, small changes in an offsprings genome, from its parents, occur randomly. Accumulation of these small changes in the genome are ultimately responsible for the formation of new species. This process is very slow and is ultimately dictated by the mutation rate u and other demographic influences. For instance, at some time in the past, there may have existed a species which we will denote as group 0. From this group 0 species, through drift, two new species may have been created, which we will denote as group 1 and 2 respectively. Hence group 1 and 2 are somehow related to each other, but are not considered the same species. Species 1 and 2 together have a common ancestor, which is species 0 (see figure 2.1). Under

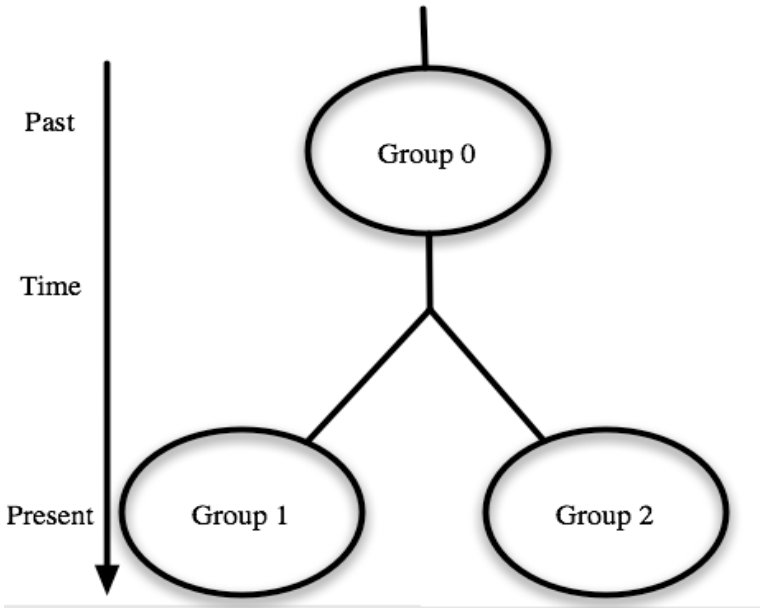

Figure 2.1: Species groups 1 and 2 both have species group 0 as a common ancestor.

the evolutionary process, two species will have existing members who through

their genealogical past share a common ancestor. It is the time of the species *divergence*, the time since the two species were created from a single species, which is of primary concern throughout this thesis. While estimation of this time is primarily of biological concern, the estimation of this time, using genomic data, is ultimately a statistical question. This divergence time we will often refer to as the time of speciation or speciation time. Other processes may affect the inference of this time, so complex models are often deployed.

Before going into depth about models of speciation, it is helpful to examine drift processes in a very simplistic framework. The next two sections explore models that were developed in the 1930s. While these models are simple, these models are illuminating in terms of motivating more complex models.

2.1.1 The Wright-Fisher model

The Wright-Fisher model (Fisher, 1930) is possibly the simplest model for genetic drift. The key assumptions involved in this model are

- finite population size,
- no overlap between generations,
- allele frequency is determined only by the drift process.

The first assumption states that the number of individuals in the population is fixed throughout time. The second assumption asserts that no individuals survive from one generation to the next. The third assumption states that no selection exists and genetic drift is neutral. While all three assumptions may seem unrealistic, they provide a starting point in which we can study genetic drift and motivate

coalescence theory.

Let there be a constant population size of N individuals. Without loss of generality, consider a scenario where there are two allele types in the population. For the description here, let us call these *red* and *black* alleles. At a given time k (measured in number of generations) in the history, let X_k be the number of *red* alleles at generation k. Under the Wright-Fisher model, the transition probability of having j *red* alleles in the $(k+1)$st generation given i *red* alleles in the kth generation is

$$P(X_{k+1} = j | X_k = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}. \tag{2.1}$$

This expression simply expresses a sampling with replacement from generation k to generation k+1. From this, we note the expected number of *red* alleles in the $(k+1)$st generation is $N \times \frac{i}{N}$. Therefore, we have that $E[X_{k+1}|X_k] = X_k$ as well as $E[|X_k|] < \infty$, so the Wright-Fisher process is a martingale. The conditional variance of X_{k+1} given X_k during each generation is $N \times \frac{i}{N}(1 - \frac{i}{N})$. The states where $X_k = 0$ or $X_k = N$ are fixed (or absorbing) states. That is, once either of these states are reached, X_k will remain in that state for an infinite amount of time.

Under the Wright-Fisher process, at each generation, the number of of descendants of a particular allele follows the binomial distribution $\text{Bin}(N, 1/N)$.
As N $\to \infty$, this distribution can be approximated by the Poisson distribution, with rate parameter $\lambda = \frac{1}{N} \times N = 1$. We let Y_{k+1} denote the number of descendents left by an individual allele in the subsequent generation. With the

approximation that $\lambda = 1$, we have

$$P(Y_{k+1} = n) = \frac{1}{n!}e^{-1}.$$

Hence the probability that any allele doesn't leave a descendent in the next generation of sampling is approximately $e^{-1} = 0.37$.

While this is a seemingly simple scenario for the drift of black and red alleles, this process can be generalized to any number of alleles and provides a reasonable basis for studying the drift process.

2.1.2 The Moran model

Like the Wright-Fisher model, the Moran model is a simple model under which the drift process can be studied. Unlike the Wright-Fisher model, where there is no overlapping of individuals between generations, the Moran model does have overlapping individuals between generations.

In the Moran model, at each generation two individuals are selected with replacement. The first individual is chosen to reproduce, while the second individual dies. All other individuals remain as they were in the previous generation. Each individual has a $\frac{1}{N}$ chance of being selected. Again, we phrase the model in terms of two alleles (*red* and *black*) and let X_k be the number of red alleles in the population at the kth generation. The Markov transition probabilities between generations follow as

$$P(X_{k+1} = j | X_k = i) = \begin{cases} \frac{i}{N}(1 - \frac{i}{N}) & \text{if } j = i + 1, \\ (1 - \frac{i}{N})\frac{i}{N} & \text{if } j = i - 1, \\ (\frac{i}{N})^2 + (1 - \frac{i}{N})^2 & \text{if } j = i, \\ 0 \quad \text{otherwise.} \end{cases} \tag{2.2}$$

It follows from these transition probabilities that

$$\begin{aligned} E[X_{k+1}|X_k = i] &= (i+1)\frac{i}{N}(1-\frac{i}{N}) + (i-1)(1-\frac{i}{N})\frac{i}{N} + i((\frac{i}{N})^2 + (1-\frac{i}{N})^2) \\ &= i. \end{aligned}$$

So, $E[X_{k+1}|X_k] = X_k$ and since X_k is bounded between 0 and N for all k, $E[|X_k|] < \infty$. Hence, the Moran process is a Martingale. It can also be shown under the Moran model that $Var(X_{k+1}|X_k = i) = 2\frac{i}{N}(1-\frac{i}{N})$.

While the variation at each step of the Moran model is less than in Wright-Fisher case, we note that at each step of the Wright-Fisher process every individual dies and is replaced at the next generation and the Moran process only allows for a maximum of one death at each step. The variance of the Moran process after two steps follows as

$$\begin{aligned} Var(X_{k+2}|X_k = i) &= E[Var(X_{k+2}|X_{k+1}, X_k = i)] + Var(E[X_{k+2}|X_{k+1}, X_k = i]) \\ &= E[2\frac{X_{k+1}}{N}(1-\frac{X_{k+1}}{N})|X_k = i] + Var(X_{k+1}|X_k = i) \\ &= \frac{2}{N}E[X_{k+1}|X_k = i] - \frac{2}{N^2}E[X_{k+1}^2|X_k = i] + Var(X_{k+1}|X_k = i) \\ &= \frac{2}{N}E[X_{k+1}|X_k = i] - \frac{2}{N^2}(Var(X_{k+1}|X_k = i) + E[X_{k+1}|X_k = i]) \\ &+ Var(X_{k+1}|X_k = i) \\ &= \frac{2}{N}E[X_{k+1}|X_k = i](1-\frac{E[X_{k+1}|X_k = i])}{N}) - \frac{2}{N^2}Var(X_{k+1}|X_k = i) \\ &+ Var(X_{k+1}|X_k = i) \\ &= Var(X_{k+1}|X_k = i) - \frac{2}{N^2}Var(X_{k+1}|X_k = i) + Var(X_{k+1}|X_k = i) \\ &= (1-\frac{2}{N^2})Var(X_{k+1}|X_k = i) + Var(X_{k+1}|X_k = i). \end{aligned}$$

For the time being, let us leave the previous two step variance formula in the current form and derive the general recursion for the M step variance. After three steps of the Moran process, the variance follows as

$$\begin{aligned} Var(X_{k+3}|X_k = i) &= E[Var(X_{k+3}|X_{k+2}, X_k = i)] + Var(E[X_{k+3}|X_{k+2}, X_k = i]) \\ &= E[2\frac{X_{k+2}}{N}(1 - \frac{X_{k+2}}{N})|X_k = i] + Var(X_{k+2}|X_k = i) \\ &= \frac{2}{N}E[X_{k+2}|X_k = i] - \frac{2}{N^2}E[X_{k+2}^2|X_k = i] + Var(X_{k+2}|X_k = i). \end{aligned}$$

Substituting the expression,

$$E[X_{k+2}^2|X_k = i] = Var(X_{k+2}|X_k = i) + E[X_{k+2}^2|X_k = i]$$

and using that $E[X_{k+M}|X_k = i] = E[X_{k+1}|X_k = i]$ (since X_k is a martingale), yields the expression

$$Var(X_{k+3}|X_k = i) = (1 - \frac{2}{N^2})Var(X_{k+2}|X_k = i) + Var(X_{k+1}|X_k = i).$$

In general we have the recursion

$$Var(X_{k+M}|X_k = i) = (1 - \frac{2}{N^2})Var(X_{k+M-1}|X_k = i) + Var(X_{k+1}|X_k = i). \quad (2.3)$$

Upon iteratively applying recursion (2.3), we obtain the solution to the M step variance of the Moran process as

$$Var(X_{k+M}|X_k = i) = \sum_{j=0}^{M-1} \left(1 - \frac{2}{N^2}\right)^j Var(X_{k+1}|X_k = i). \quad (2.4)$$

In the Moran process, each individual has a $1/N$ chance of dying at each step, so each individual is expected to live N steps before death. Hence, it is reasonable

to equate a Moran generation to N steps of the Moran process. With this, we have $E[X_{k+N}|X_k = i] = i$ and from equation(2.4), the variance follows as

$$Var(X_{k+N}|X_k = i) = \sum_{j=0}^{N-1} \left(1 - \frac{2}{N^2}\right)^j Var(X_{k+1}|X_k = i).$$

For N large, we have that $\left(1 - \frac{2}{N}\right)^j \approx 1$, so

$$\begin{aligned} Var(X_{k+N}|X_k = i) &\approx N \times Var(X_{k+1}|X_k = i) \\ &= 2N \times \frac{i}{N}(1 - \frac{i}{N}). \end{aligned}$$

This is an interesting result, since for N generations of the the Moran process, the variance is approximately twice as large as in the Wright-Fisher process.

It is of course true that genetic variation occurs through the generational reproduction of children, which always happens in a forward direction in time. While the realization of this forward process is the full genealogical history, in practice, we only get to observe the present state of allele frequencies and must infer the history from the sample (the present state of alleles).

Forward models, such as the Wright-Fisher process or the Moran process can be used to construct probability distributions about the unknown parameters, such as N in the process. Unfortunately, closed form probability distributions resulting from complex forward processes seldom exist, so likelihood functions or probability distributions for the unknown distributions are difficult to analyze via the forward process framework. With respect to computation, in order to construct probability distributions using forward processes, one might, for a given set of unknown parameters, run the process until the pattern of variation observed in a

given data set is observed. The relative frequency of times the observed data is simulated would be an estimate of the probability of the data. Performing this task for a 1-dimensional might be feasible, but over a high dimensional data set, the computation time required is a limiting factor.

The Moran and Wright-Fisher models both represent relatively simple drift processes with straightforward stochastic realizations. This is in part due to the fact that both processes, as described, don't include a mutational model or mechanism in which allele types can change from generation to generation after production. For now, we will proceed without mutation, but the issue of mutational processes will be the topic of subsequent discussions within this thesis.

Since mutation has been neglected in both the Wright-Fisher and Moran models, at some point in time only one allelic type will exist in the sample. Once this occurs, the process will be fixed in that allelic state. This is referred to as an absorbing state of the process. It can be shown by calculating the stationary distribution for both the Wright-Fisher and Moran processes that

$$\begin{aligned} \lim_{k\to\infty} \mathrm{P}(X_k = N | X_0 = i) &= \frac{i}{N} \\ \lim_{k\to\infty} \mathrm{P}(X_k = 0 | X_0 = i) &= 1 - \frac{i}{N}. \end{aligned}$$

That is, the probability of fixation is equal to the proportion of *red* (or *black*) alleles in the initial sample. This result could have also been rationalized intuitively.

2.2 A backwards perspective

We will set our focus now on models that do not mimic the natural time progression (forward). These models will start at the data level and run backwards until some finishing state of the data.

2.2.1 Coalescence models

Before directly addressing coalescence models, or simply the coalescent, consider figure 2.3, which represents a realization under the Wright-Fisher process.

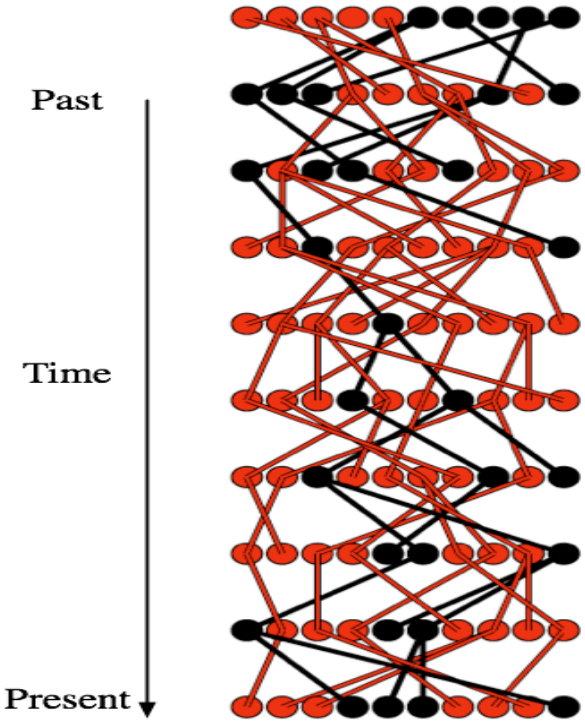

Figure 2.2: A realization of genetic drift under the Wright-Fisher process.

The genealogy from a sample of alleles, constructed through the forward process, can be readily viewed through a backward perspective. Branches which split

in the forward direction are said to coalesce through a backwards trajectory of the sample. In a forward direction, lineages can be terminated before the present day state (bottom of the genealogy) by not being selected in the generational transition. However, at the present state, going backwards, all genealogies proceed without termination. Figure 2.3 shows a realization of the Wright-Fisher process (left and central panels), and a backwards tracing of two individuals in the present sample (rightmost panel). The point in which the two samples find their most recent common ancestor is the point of coalescence.

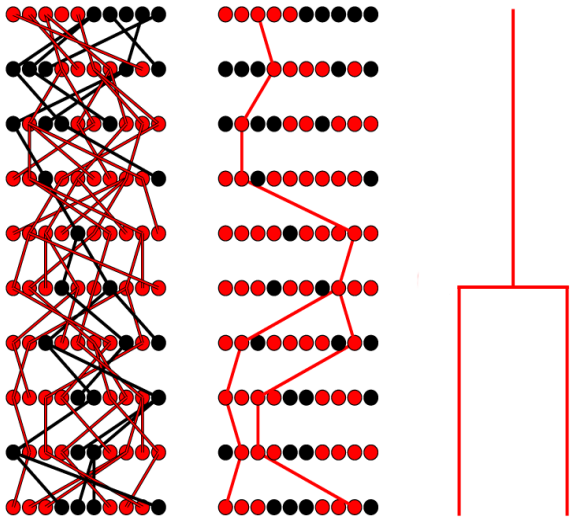

Figure 2.3: A realization of genetic drift under the Wright-Fisher process with a backward tracing of two samples in the present sample.

We notice that in the present day state, the alleles have a genealogy which is represented by the backwards trajectory from all the alleles at the bottom state of the ancestry.

While tracing back two individuals in the sample is the fundamental principal in a coalescence model, this principal can be applied to any arbitrary number of individuals. Figure 2.4 illustrates the back tracing (coalescing) of three individuals in the sample (present state), as produced by the Wright-Fisher model. The left panel in figure 2.4 represents a realization of the Wright-Fisher process, while the central panel shows a back tracing of three individuals in the sample. The right most panel shows an untangled version of the genealogy, for the three chosen individuals in the sample.

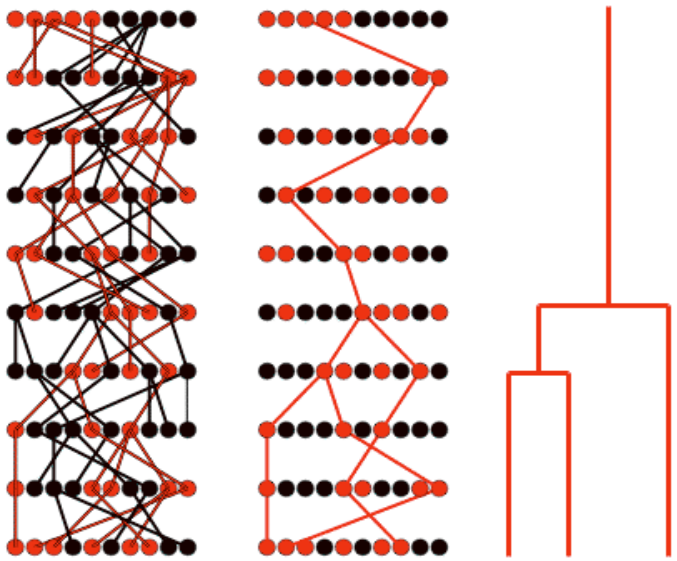

Figure 2.4: A realization of genetic drift under the Wright-Fisher process with a backward tracing of three samples in the present sample.

The following derivation is a standard result in coalescence theory and can be found in Hudson (1990) or Felsenstein (2003). If we do not observe the genealogy, then each allele has probability $\frac{1}{N}$ of coming from any particular parent in the previous generation. Hence, the probability that a given pair of alleles do not share the same parents in the previous generation is $p_2 = 1 - \frac{1}{N}$. The probability

that a given third allele has a distinct parent as the previous two alleles (given that they have distinct parents) in the previous generation is given by $1 - \frac{2}{N}$. Therefore the probability that all three of these alleles have different parents in the previous generation is given by

$$p_3 = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right).$$

From this, we see that the probability of k genes all having distinct parents in the previous generation is given by

$$p_k = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\ldots\left(1 - \frac{k-1}{N}\right).$$

Expanding this expression yields

$$\begin{aligned} p_k &= 1 - \frac{1}{N}\sum_{i=1}^{k-1} i + O\left(\frac{1}{N^2}\right) \\ &= 1 - \frac{1}{N}\binom{k}{2} + O\left(\frac{1}{N^2}\right), \end{aligned}$$

where $O\left(\frac{1}{N^2}\right)$ is a sum with terms involving $\frac{1}{N^j}$ with $j \geq 2$. These terms represent the probability contribution from three or more alleles sharing the same parent in the previous generation. For large N, we obtain the approximation for the probability of all k alleles having distinct parents in the previous generation as

$$p_k \approx 1 - \frac{\binom{k}{2}}{N}.$$

This is simply a statement that when N is large, the probability that three or more descendants come from the same parent in the previous generation is vanishingly

small. From this expression for p_k, the probability that two alleles share a parent in the previous generation is given by

$$1 - p_k = \frac{\binom{k}{2}}{N}.$$

This probability simply states, the probability that any two alleles have probability $\binom{k}{2}/N$ of coalescing in the previous generation. This could have been intuitively derived by noting that for a given set of alleles, the probability they share the same parent in the previous generation is $\frac{1}{N}$. Since there are $\binom{k}{2}$ ways of selecting two alleles, we obtain the probability of $\binom{k}{2}/N$ of coalescence in the previous generation. From this, we see that the number of generations until two individuals coalesce has a geometric distribution with parameter $p = \frac{\binom{k}{2}}{N}$. If we denote G as the number of generations until coalescence, then the probability of waiting g generations for coalescence is given by

$$P(G = g) = \left(1 - \frac{\binom{k}{2}}{N}\right)^{g-1} \frac{\binom{k}{2}}{N}.$$

The expected number of generations until a coalescence event is given by

$$E[G] = \frac{N}{\binom{k}{2}}. \tag{2.5}$$

Instead of using a discrete generational process, it is convenient to model the coalescence process through a continuous time model. The exponential distribution provides a continuous time approximation to the geometric distribution. In terms of time (instead of generations), the time to coalescence is approximated by an exponential distribution, so that when there are k individuals in the sample, the

time to coalescence is modeled by

$$P(t_k = t) = \exp(-t\frac{\binom{k}{2}}{N}), \tag{2.6}$$

where t_k is the time to coalescence when there are k individuals in the sample. Notice that

$$E[t_k] = \frac{N}{\binom{k}{2}},$$

which is equal to the expected number of generations required for coalescence under the discrete model.

We define the coalescent to be the complete set of ordered coalescing events, which begins with the sample and ends once all the individuals have coalesced into a single entity, which is termed the Most Recent Common Ancestor (MRCA) of the sample. We may wish to know how long this actually takes to occur. That is, we might want to know how long we expect to wait until all lineages in the sample have coalesced. This is simply derived by summing over all levels of the coalescent. Letting T_{MRCA} denote the time required to reach the MRCA, we have

$$T_{MRCA} = \sum_{j=2}^{k} t_k. \tag{2.7}$$

By taking the expectation of both sides of equation (2.7), we have

$$\begin{aligned}
E[T_{MRCA}] &= E[\sum_{j=2}^{k} t_j] \\
&= \sum_{j=2}^{k} E[t_j] \\
&= \frac{N}{\binom{k}{2}} + \frac{N}{\binom{k-1}{2}} + \cdots + \frac{N}{\binom{2}{2}} \\
&= 2N\left[\left(\frac{1}{k-1} - \frac{1}{k}\right) + \left(\frac{1}{k-2} - \frac{1}{k-1}\right) + \cdots + \left(\frac{1}{1} - \frac{1}{2}\right)\right] \\
&= 2N\left(1 - \frac{1}{k}\right).
\end{aligned}$$

We see that when k is large, the expected length of the tree is approximately $2N$. Also, note, that when $k = 2$, that is when there are two individuals waiting to coalesce, the expected waiting time is N. Therefore, the waiting time for the last two individuals to coalesce nearly consumes half of the waiting time over the

whole coalescence event.

Calculating the total expected length of a genealogy with k individuals at the base of the genealogy illustrates the power of a coalescence analysis. It is this backward perspective that enables us to easily incorporate all the boundary parameters of the stochastic realization, that is incorporation of the sample itself. While the forward process is ultimately responsible for producing the sample, inference is often times simpler under a coalescence based analysis.

2.3 Mutation

When we include mutations into the process, a coalescence based approach becomes even more attractive since genealogies compatible with the sample are easy to construct and mutations can easily be superimposed on the genealogy constructed via coalescence.

When mutations are allowed to occur in the drift process, it is not necessarily the case that a descendant will inherit the allelic state of its parent. Consider a realization of the Wright-Fisher in which mutations arise per individual, per generation, with probability u (see figure 2.5) Inclusion of mutation into the forward process now means that the boundary states 0 and N are not absorbing. The coalescence process is easily modified to incorporate mutation into the drift process.

It is not exactly possible to measure incremental units of time by the examination of gene samples alone, since we do not know what the population size N is and we don't know the initial state of the population. However, measuring the diver-

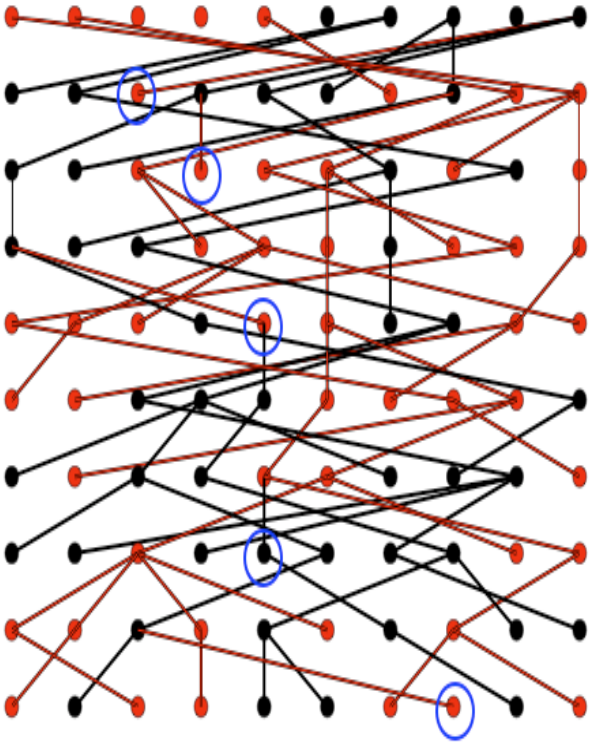

Figure 2.5: A realization of genetic drift under the Wright-Fisher process with mutation. Individuals who have experienced a mutation have been highlighted with blue circles.

gence in genetic structure is possible by incorporation of a known mutation rate u.

We assume that mutations arise independently, as they would under neutral evolutions. That is, the existence of a mutation will not inspire new mutations (as is the case with deleterious mutations which effect fitness). Since u is the probability that a gene experiences a mutation, the number of mutations D_t a particular gene will experience in t generations has (approximately) the Poisson distribution

$$P(D_t = d) = \exp(-tu)\frac{(tu)^d}{d!}. \tag{2.8}$$

If we observe to random individuals from a population, we may wish to know how many mutational differences they have between them, for a given gene. Again, we let the mutational rate for this gene be u and denote the number of mutational differences between the individuals as π. For convience, we assume the infinite sites model, where all mutations are observed in the sample. Since these genes ultimately came from some common ancestor t time units ago, we have

$$E[\pi] = 2 \times E[t] \times u. \tag{2.9}$$

The factor of 2 in equation (2.9) exists since mutations will occur on both lineages of the two individuals. From equation (2.5), we obtain

$$\begin{aligned} E[\pi] &= 2 \times E[t] \times u \\ &= 2Nu \end{aligned}$$

since $E[t] = N$.

Traditionally, we denote

$$\theta = 2Nu$$

Hence the parameter θ represents the expected number of mutational differences between sequences. Also, θ can be seen as a scaled effective population size N, where the scaling is by a multiple of the mutation rate u. This mutation parameter u is often described as the *molecular clock* or *mutation clock* since it can give us an indication of time (it is measured in units of mutations/time), given the number of mutations observed in the sample. Typical mutation rate values for eukaryotes (i.e. plants, animals, fungi) are in the range $10^{-4} - 10^{-6}$ (Kumar and Subramanian, 2002).

In a coalescence model, it is not possible to estimate population parameters such as N separately from u. In subsequent chapters we will be interested in estimating *clock* scaled parameters like θ for a given data set. Although, all parameters within a coalescence framework will only be estimable through a product with the mutation rate, the mutation rate can often be estimated through other experimental means. We will not go into detail about this process, but rather take it for granted that the mutation rate u is known.

2.3.1 Distribution of mutations

In order to study a given data set, a model must be specified for how mutations are accumulated during the genealogical history of the sample. One such model is to let the mutations occur under a similar distribution as coalescence events. That is, we might wait for an exponential amount of time before a mutation event is observed. We let X_u denote this time. For two diverging individuals, we let D_{2u} denote the number of mutations that arise during a fixed amount of time t. It follows from the exponential waiting time assumption that D_{2u} has a Poisson distribution. Since we are considering two individual genes that are diverging from a common ancestor, the rate of mutation for both sequences is $2u$. Also, since time in a forward model is the same as the time under a backward model, we have that the number of mutations accumulated between two coalescing individuals follows the Poisson distribution

$$P(D_{2u} = d|u, N, t) = e^{-2tu}\frac{(2tu)^d}{d!}. \tag{2.10}$$

We have let continuous time be denoted as t, however, it is convenient to scale time by N generations as

$$\tau = \frac{t}{N}. \tag{2.11}$$

With this we can rewrite equation (2.10) as

$$\begin{aligned} P(D_{2u} = d|u, N, t) &= e^{-2tu}\frac{(2tu)^d}{d!} \\ &= e^{-2Nu\frac{t}{N}}\frac{(2Nu\frac{t}{N})^d}{d!} \\ &= e^{-\theta\tau}\frac{(\theta\tau)^d}{d!} \\ &= P(D_{2u} = d|\theta, \tau). \end{aligned}$$

The distribution gives the probability of observed d pairwise mutations between two sequences given it takes τ time units to coalesce. We wish to know the distribution of mutations not dependent on this time (since we don't actually know it). Recall (from equation (2.6)) that the time required to coalesce two genes is has the exponential density

$$p(t_k = t) = \frac{1}{N}e^{-t/N}.$$

Under the transformation in (2.11), we have $p(\tau) = e^{-\tau}$. From this we derive the probability of observing d pairwise mutations between two samples follows as

$$\begin{aligned} \mathrm{P}(D_{2u} = d|\theta) &= \int_0^\infty \mathrm{P}(d|\theta, \tau)\mathrm{p}(\tau)d\tau \\ &= \int_0^\infty \frac{e^{-\theta\tau}(\theta\tau)^d}{d!}e^{-\tau}d\tau \\ &= \int_0^\infty \frac{e^{-\tau(\theta+1)}(\theta\tau)^d}{d!}d\tau \end{aligned} \tag{2.12}$$

Recognizing the integrand in equation (2.12) as the kernel of the Gamma distribution:

$$g(x|\alpha,\beta) = x^{\alpha-1}\frac{\beta^{\alpha}}{\Gamma(\alpha)}e^{-\beta x},$$

and noting that $\Gamma(d+1) = d!$, we can rewrite equation (2.12) as

$$\begin{aligned}
\mathrm{P}(D_{2u} = d|\theta) &= \int_0^{\infty} \frac{e^{-\tau(\theta+1)}(\theta\tau)^d}{d!}d\tau \\
&= \int_0^{\infty} \tau^d \frac{(\theta+1)^{d+1}}{(\theta+1)^{d+1}}\frac{\theta^d}{\Gamma(d+1)}e^{-\tau(\theta+1)}d\tau \\
&= \theta^d \frac{1}{(\theta+1)^{d+1}}\int_0^{\infty} \tau^d \frac{(\theta+1)^{d+1}}{\Gamma(d+1)}e^{-\tau(\theta+1)}d\tau \qquad (2.13)
\end{aligned}$$

It follows that

$$\int_0^{\infty} \tau^d \frac{(\theta+1)^{d+1}}{\Gamma(d+1)}e^{-\tau(\theta+1)}d\tau = 1$$

since the integrand has a Gamma distribution with parameters $(\alpha,\beta) = (d+1, \theta+1)$. Therefore we have

$$\begin{aligned}
\mathrm{P}(D_{2u} = d|\theta) &= \theta^d \frac{1}{(\theta+1)^{d+1}} \\
&= \left(\frac{\theta}{\theta+1}\right)^d \frac{1}{\theta+1}. \qquad (2.14)
\end{aligned}$$

Hence the distribution of mutations (only conditional on θ) has a geometric distribution.

This result is easily generalize to the case when there are k individuals eligible for coalescence. Letting D_{ku} be the number of mutations that accumulate in

the interval separating the k^{th} and $k-1^{st}$ levels of the coalescent is

$$\begin{aligned} \mathrm{P}(D_{ku}|\theta,k) &= \left(\frac{ku}{ku+\binom{k}{2}/N}\right)^d \frac{\binom{k}{2}/N}{ku+\binom{k}{2}/N} \\ &= \left(\theta+\frac{\theta}{(k-1)}\right)^d \frac{k-1}{\theta+(k-1)}. \end{aligned}$$

From this, the expected number of mutations accumulating between the k^{th} and $k-1^{st}$ levels is

$$E[D_{ku}|k] = \frac{\theta+(k-1)}{k-1}.$$

Letting M_{tot} be the total number of mutations accumulated over the whole coalescent, when we start with j individual genes, we have in expectation

$$\begin{aligned} E[M_{tot}] &= \sum_{k=j}^{2} \frac{\theta+(k-1)}{k-1} \\ &= \sum_{1}^{j-1} \frac{\theta+k}{k} \\ &= \sum_{1}^{j-1} \left(\frac{\theta}{k}+1\right) \\ &= \left(\theta\sum_{1}^{j-1}\frac{1}{k}\right)+j-1. \end{aligned}$$

For large j, we obtain the approximation

$$\begin{aligned} E[M_{tot}] &= \left(\theta\sum_{1}^{j-1}\frac{1}{k}\right)+j-1 \\ &\approx \theta(\gamma+\log(j-1))+j-1, \end{aligned}$$

where $\gamma \approx 0.577$ is Euler's constant.

2.4 The Ewens sampling formula

In this section we will introduce the Ewens Sampling Formula (ESF), which is a sampling distribution for the number of alleles and their frequencies in a sample of n genes. For this, we assume that there are an infinite number of alleles in the population, which is known as the infinite alleles assumption in population biology. The underlying assumption of infinite alleles is that *any new mutation will result in a new allele* (resulting in an infinite number of possible alleles).

Under this notion of infinite alleles, the ESF yields the probability that a sample of n genes will contain k allelic types and that there will be $A = (a_1, a_2, \ldots, a_n)$ alleles represented $1, 2, \ldots, n$ times in the sample. As an illustration, consider figure 2.6. Figure 2.6 illustrates a sample of 10 alleles. In this sample, there are

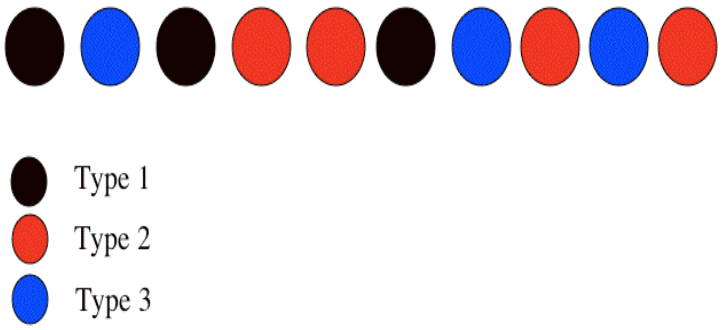

Figure 2.6: A sample of 10 alleles.

3 allelic types, where $A = (0, 0, 2, 1, 0, 0, 0, 0, 0, 0)$. Notice that we must have

$$\sum_{i=1}^{n} i a_i = n$$

and,

$$\sum_{i=1}^{n} a_i = k.$$

Ewens (Ewens, 1972) derived the ESF as

$$P(k, a_1, a_2, \ldots, a_n) = \frac{n!\theta^k}{\theta_{(n)}} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!}, \tag{2.15}$$

where $\theta_{(n)} = \theta \times (\theta + 1) \times \cdots \times (\theta + n - 1)$.

Under this description of the ESF, $\theta = 4Nu$, which is twice the mutation rate in which we've been accustomed to using. This simple modification reflects the description of a diploid rather than a haploid population. Since ESF is actually the joint probability distribution of observing k allelic types in a sample of n genes with the assignment of alleles given by $(a_1, a_2, \ldots, a_n)$, we will also refer to the Ewens sampling formula as Ewens distribution. Karlin and McGregor (1972) added an addendum to Ewens' original proof to help illuminate some of the more difficult theoretical issues in the original paper.

2.4.1 ESF derivation

While the conception of the ESF predates the coalescence, the ESF is much easier to derive from a coalescence argument. The original proof for equation (2.15) is quite difficult, and even subsequent simplified re-derivations of the Ewens distribution tend to be complex. We will employ a derivation of Ewens sampling distribution as described in Griffiths and Lessard (2005).

The number of ways in which k allelic types can be distributed in a sample of

n genes is

$$\frac{n!}{\prod_{i=1}^{k} n_i},$$

where the n_i's are the frequency of each allele. For our example, $n_1 = 3$, $n_2 = 4$, $n_3 = 3$. Since the k allelic types in the sample are not labeled, these can also be permuted. For example, in figure 2.6, we have three allele types where type 1 and type 2 alleles both 3 representatives in the sample. Hence these allele types can be permuted in $a_3!$ ways without changing the overall configuration of the sample. In general there are a total of

$$\prod_{i=1}^{n} a_i!$$

of these permutations. Hence there is a total of

$$\frac{n!}{\prod_{i=1}^{k} n_i! \prod_{i=1}^{n} a_i!}$$

unlabeled configurations of the n genes.

Now we must account for how the variation in the sample arose under the infinite allele model in the first place. This is most easily accounted for through a coalescence perspective. At this point, we employ the infinite alleles assumption. That is, only a single mutation is responsible for the formation of the novel allele. This means that in a backwards (coalescence) perspective, the allele must ultimately disappear due to mutation.

Figure 2.7 shows that each allele type in a sample of 10 alleles was generated by a single mutation. From a backwards coalescence perspective, this mutation must be the last event to occur before coalescing with a different allele type.

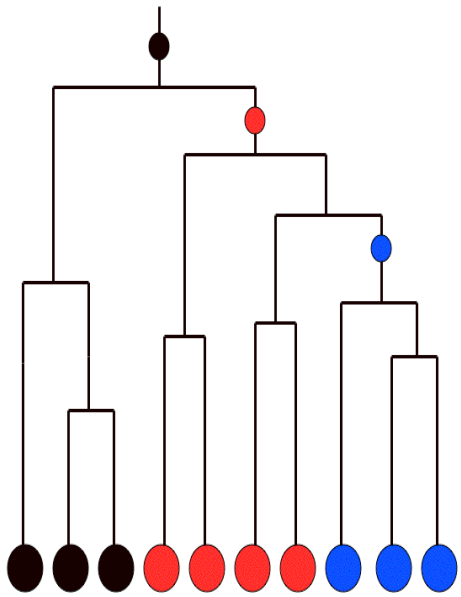

Figure 2.7: A possible genealogy responsible for the resulting allele types observed in a sample of 10 genes.

We let transitions refer to a change in state to the sample, either by mutation or a coalescence. When there are i genes left in the sample ,the overall rate of transition is

$$iu + \frac{\binom{i}{2}}{N}.$$

The rate at which a single allele disappears due to mutation is u, so the probability that a *given* allele disappears due to mutation, when there are i individuals left, is

$$\frac{u}{iu + \frac{\binom{i}{2}}{N}} = \frac{\theta}{i(\theta + i - 1)}.$$

When there are j individuals of the same allelic type left in the sample (and i total individuals left), there are $\binom{j}{2}$ possible ways of choosing any of these j individuals to coalesce. Therefore the rate at which a *particular* individual coalesces with another individual is $\frac{1}{j}\binom{j}{2}$. It follows that the probability that an individual is lost due to coalescence, where there are j individuals of that type is

$$\frac{\frac{1}{j}\binom{j}{2}}{iu + \frac{\binom{i}{2}}{N}} = \frac{j-1}{i(\theta + i - 1)}.$$

The joint probability of a given ordered history of mutational losses and coalescence events follows as

$$\begin{aligned} & \frac{(n_1-1)!\ldots(n_k-1)!\theta^k}{1\theta \times 2(\theta+1) \times \cdots \times n(\theta+n-1)} \\ = \quad & \frac{(n_1-1)!\ldots(n_k-1)!\theta^k}{n!\theta_{(n)}}, \end{aligned}$$

which is the product of coalescence probabilities and mutation probabilities for a particular genealogical history or the sample.

Since there are n events in the coalescing history, there are $n!$ possible histories. Hence the overall total probability of the sample follows as

$$\begin{aligned} P(k, a_1, a_2, \ldots, a_n) &= n! \frac{n!}{\prod_{i=1}^{k} n_i! \prod_{i=1}^{n} a_i!} \frac{(n_1-1)!\ldots(n_k-1)!\theta^k}{n!\theta_{(n)}} \\ &= \frac{n!}{\prod_{i=1}^{k} n_i \prod_{i=1}^{n} a_i!} \frac{\theta^k}{\theta_{(n)}} \end{aligned}$$

.

Noting that

$$\prod_{i=1}^{k} n_i = \prod_{i=1}^{n} i^{a_i},$$

we arrive at $P(k, a_1, a_2, \ldots, a_n) = \frac{n!\theta^k}{\theta_{(n)}} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!}$, which is the ESF!

This derivation of Ewens distribution makes use of the coalescent and tremendously simplifies the mathematics over other derivations which employ complicated combinatorial analyses, recursions, or Poisson-Dirichlet process insights.

Chapter 3

Decoupled Parameter Space Models

A coalescence process can be considered a pure death process in which the number of elements in the state space decreases due to the merging of two elements at a time. This merging process is referred to as coalescing. This process starts from the number of initial samples in the data set, and proceeds until all the individuals have coalesced into a common element. This common element is what we shall refer to as the most recent common ancestor (MRCA). The stochastic realization (path) which describes the pattern of coalescences, from the individuals in the sample to the MRCA, represents the gene genealogy of the studied data set. This gene genealogy we will refer synonymously to as a history, demographic, or simply a tree due to its bifurcating nature.

In general, inference can include anything on the tree, such as cumulative branch lengths from the sample, which might represent the time since an event occurred, as well as include any parameters which drive the stochastic process. In this chapter we will introduce a set of coalescence models, in which all time inferences are restricted to the parameters that drive the process, and not the stochastic real-

izations themselves. We will denote $\boldsymbol{\theta}$ as the population parameters that are of primary interest which drive the stochastic process, and T the evolutionary tree which will be the stochastic outcomes of the process. We impose that if event times in the genealogy are wanted, they must be included in the set $\boldsymbol{\theta}$ and not explicitly on the sample path T. The feature that event times are contained in the set $\boldsymbol{\theta}$, rather than T is what we call a decoupling of the system. Hence T, in our view, serves completely as a latent process and merely acts as a modeling device.

Models of coalescence generally share the exponential waiting time till event framework and exploit the fact that modeling the backward process greatly simplifies computational demands and clarifies theoretical insights of the drift process. Key to the models studied in this chapter is that the parameters of interest and stochastic realizations are decoupled. In mathematical terms, by a decoupling, we mean we can easily seperate the joint posterior distribution

$$P(T, \boldsymbol{\theta}|D)$$

into the components

$$P(T|\boldsymbol{\theta}, D) \qquad \text{and} \qquad P(\boldsymbol{\theta}|T, D).$$

The notion of branch length will not be included in the description of T, but rather T will simply be an ordering of events. The parameter set $\boldsymbol{\theta}$ will serve as functions of the rate parameters in sampling distributions that model waiting times, so the sampling distribution of the particular times of interest may be inferred through marginalization of T.

There are many reasons for requiring this decoupling. One reason being that

it lends it self to a simpler extraction of the information since we really only desire the distribution $P(\theta|D)$ for inference. Also, being that $\boldsymbol{\theta}$ only contains numerical quantities, whereas T is a high dimensional and oddly topological construction, it makes sense to separate the numerical quantities of interest from the non-numerical elements. In later chapters, we will build sampling methodologies which are much more efficient, from a computational perspective, due to this decoupled structure.

3.1 The Data

Before introducing a thorough description of the coalescence models for the particular demographic parameters that we wish to study, let us describe the observed data and the summary statistics that we will explore. Our data comes from nucleotide sequence data, which for this chapter we will consider to be from a single locus. That is, from multiple individuals, nucleotide sequence data is extracted which will come from a single common region of the genome and summary statistics will be formed from the sequence data. A nucleotide sequence (or DNA sequence) is a succession of the 4 bases A,G,C and T which stand for adenine, guanine, cytosine and thymine respectively. A sequence can consist of any number of bases. An example of a nucleotide sequence is

$$GGGAACCTAAGACCTAGATCAAGGCCA.$$

3.1.1 Outgroups

Before describing the summarization of the sequence data, let us consider the notion of an *outgroup* species. For any two species, say species 1 and species 2,

a third species, say species 3, is called an *outgroup* if it is more distantly related to species 1 and species 2 than they are to each other. In terms of a phylogeny, the topology of species 1, species 2 and the outgroup species can be seen in figure 3.1. The concept of an outgroup is necessary for classifying different mutations

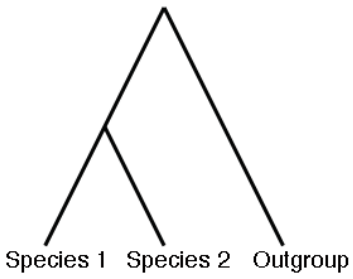

Figure 3.1: A topological representation of an outgroup species.

in the data set and determination of the ancestry of the mutation. Consider for example the two observed nucleotide sequences

$$AAGTTC$$

$$AACTTC$$

which come from two species, species 1 and species 2 respectively. The two sequences are identical in all but the third position. This third position, since it differs in both sequences, is known as a polymorphic site, polymorphism or a segregating site. We know that at least one mutation occurred in the evolutionary history at site three. Under the infinite sites model, we assume exactly one mutation occurs per site in the entire history of the sample. With the assumption of infinite sites, we can use information from an outgroup sequence for identifying

which sequence is mutated and which is ancestral. For example, if the outgroup sequence was

$$AACTTC,$$

we would know the mutation occurred somewhere in the history of species 1, but not in the history of species 2. This evolutionary history is depicted in figure 3.2.

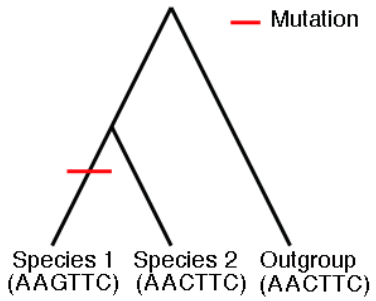

Figure 3.2: The outgroup species identifies which species group develops the mutation.

3.1.2 Segregating Sites And Watterson's Estimator

A common parameter of interest in population genetics is $\theta \propto Nu$, where u is the mutation rate and N is the effective population size for a single population. This parameter, θ, is essentially of interest for describing the effective size of a population but, as we shall see, can only be estimated relative to the mutation rate u. Segregating sites from a sample of DNA data are base positions where the sequences differ. Watterson (Watterson, 1975) contributed an estimator for θ in which knowing the total count of segregating sites (total number of base positions

that show differences) is sufficient. Although, Watterson did not originally use a coalescence approach in constructing this estimator, we will motivate his estimator via coalescence.

Consider for example being at a position on the evolutionary tree, in which there are k branches and we are waiting to coalesce to $k-1$ branches. The mutation rate on any one of these branches is u, so the total mutation rate for all k branches is uk. From chapter 2, the waiting time to coalesce from k branches to $k-1$ branches is

$$\frac{2N}{k \times (k-1)}.$$

From this, we have the expected number of accumulated mutations on the level of the tree with k branches as

$$\frac{2N}{k \times (k-1)} \times uk = \frac{\theta}{(k-1)},$$

where we have taken to be $\theta = 2Nu$. To determine the total number of mutations (segregating sites) that occur in the samples (DNA sequences) evolutionary past, we must count the number of mutations that occur on all levels of the tree. Hence the expected number of segregating sites follows as

$$E[S] = \theta \sum_{k=2}^{L} \frac{1}{k-1} = \theta \sum_{k=1}^{L-1} \frac{1}{k},$$

where L is the total number of samples. From this, the method of moments (MOM) estimator for θ follows as

$$\hat{\theta} = \hat{S} / \sum_{k=1}^{L-1} \frac{1}{k},$$

where $\hat{S}$ is the total number of segregating sites observed in the L samples.

3.1.3 Classification Of Segregating Sites

Wakeley and Hey (1997) extend the idea of counting segretating sites to consider estimation of a larger set of population parameters, which model the demographic history of two species groups, by considering the number of sites which segregate in both groups, segregating in a single group and those with fixed differences (see figure 3.3).

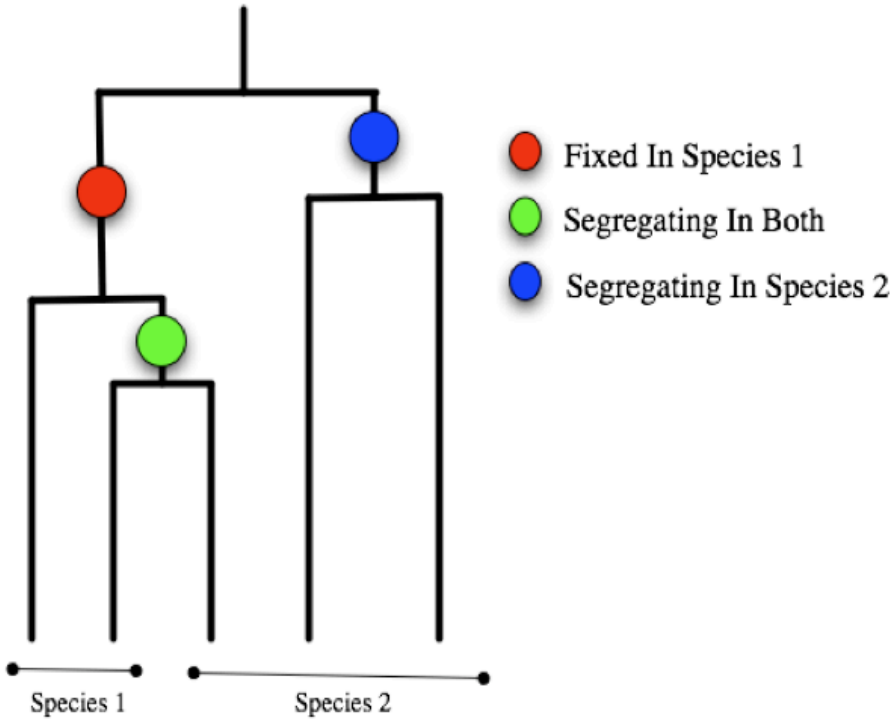

Figure 3.3: The Wakley Hey classification of segregating sites.

We will further extend this concept of counting segregating sites based on the classification of mutations, where derived mutations can be either absent (not occurring), segregating (occurring in at least one but not all), or fixed (occurring

in all). Mutations will be jointly classified in a joint two species configuration based on the fixed, absent or segregating types. This joint classification defines seven types since mutations are restricted from being fixed or absent in both groups simultaneously. Table 3.1 shows the classification of the seven types of segregating sites $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5, n_6, n_7)$.

Count	Group 1	Group 2
n_1	segregating	absent
n_2	fixed	absent
n_3	absent	segregating
n_4	absent	fixed
n_5	segregating	segregating
n_6	fixed	segregating
n_7	segregating	fixed

Table 3.1: Counts of segregating sites provide multiple summary statistic from nucleotide sequences.

An example of the data we might observe is depicted in table 3.2.

Group	Sequence Data
out	AAAGGCTTAACCGGTTATTA
1	AGAGGCTTAATCGGTTATTT
1	AGAGGCTTAATCGGTTATTA
1	AGAGGCTTAACCGGTTATTA
1	AGAGGCTTAACCGGTTATTA
2	AAAGACTTAACCGGTTAATA
2	AAAGACTTAACCGGTTATTA
2	AAAGACTTAACCGGTTATTA

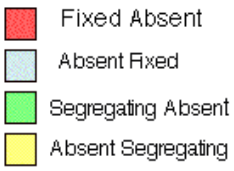

Table 3.2: An example of the observed data.

We see in table 3.2, there are 4 types of mutations observed in the data set, these are of type fixed/absent, absent/fixed, segregating/absent and absent/segregating. These 4 mutational types have occurred with frequencies $\{1, 1, 2, 1\}$ respectively, and the total number of segregating sites ($\hat{S}$) is 5.

Unlike Watterson's estimator for θ, where only the total number of segregating sites is sufficient, classifying the counts of segregating sites yields information about the genealogical history by determining classes of branches that must be present in the gene tree.

For example, from our running example, a possible topology of the tree with mutational placements is shown in figure 3.4.

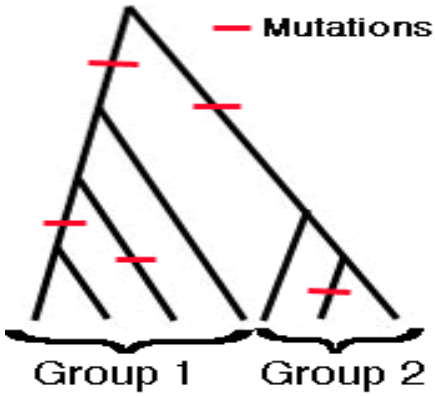

Figure 3.4: An example of the tree structure that is induced by the classification of mutations.

Notice for example, the highest mutation in the tree in figure 3.4 corresponds to a branch that has fixed descendents in species 1, and no descendent in species 2. So, this mutation is of type n_2 and describes the fixed/absent mutation seen in table 3.2. Hence, the joint classification of mutations, as defined by table 3.1 yields information about the gene tree. In subsequent chapters, we will demonstrate how good this summarization of the data is by comparing the inference under models conditioned on this data to models conditioned on the fully un-summarized data.

3.1.4 State Space

We define the level of a tree by the number of branches that exist between consecutively ordered coalescence events. One a given tree, level l denotes the section of the tree where l lineages exist (see figure 3.5).

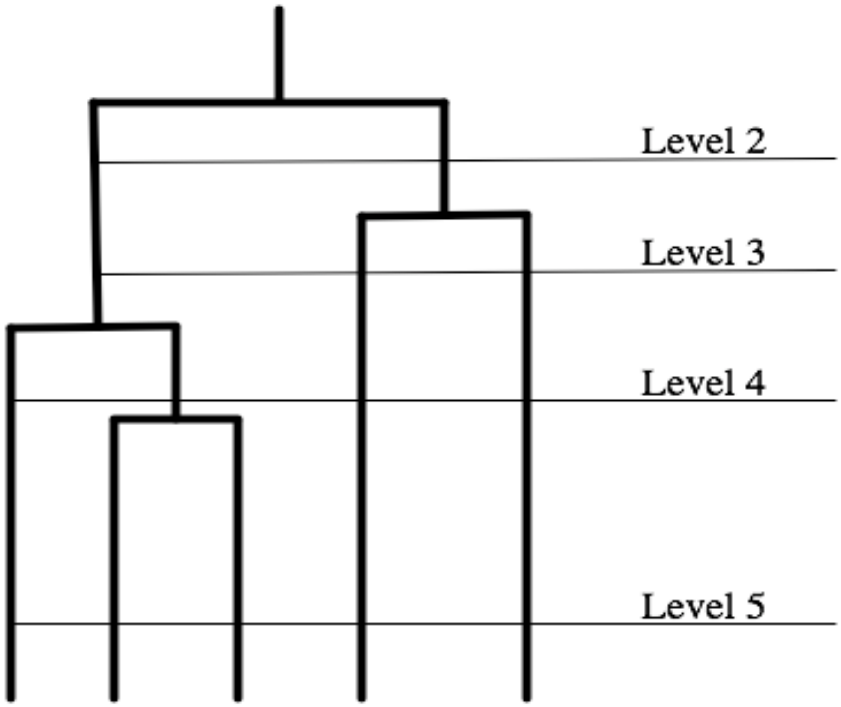

Figure 3.5: Levels of the tree are denoted by the number of branches that exist between consecutive coalescence events.

Every branch of the gene genealogy is classified by lineage according to its species membership. That is we note to what group(s) its descendants belong. A type 1 lineage has descendants in only the species 1 group, and a type 2 lineage has descendants in only the species 2 group. Type 3 lineages have descendants in both groups.

On level l of the geneology, where there are l genes still available for coalescence,

we denote the state space as

$$\boldsymbol{S}_l = (l_{1,1}, l_{1,2}, l_{1,3}|l_{2,1}, l_{2,2}, l_{2,3}|l_{0,1}, l_{0,2}, l_{0,3}),$$

where $l_{i,j}$ denotes the number of lineages, on level l, that belong to the species i group and are of type j.

It follows that $l = \sum_{i=1}^{3} \sum_{j=0}^{2} l_{i,j}$ and $(l_{1,j} + l_{2,j}) \times l_{0,j} = 0$, since we are either in the post or prespeciation phase.

The model state space reflects the types of branches on the genealogical tree at each level l. A full sequence of state spaces $\boldsymbol{S}_l$, for $l = (n, \ldots, 2)$, where n is the number of individuals in the initial sample, denotes a tree structure. An example of the a tree structure and its (partial) state space representation is illustrated in figure 3.6.

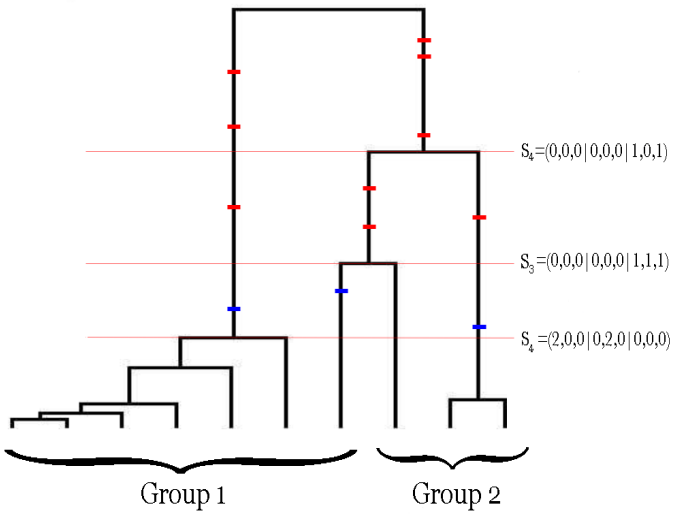

Figure 3.6: A genealogical representation of the state space $\boldsymbol{S}_l$ for each level of the tree. Notice that between level 4 and level 3 there was a change in population structure, induced by a speciation event which transitioned the state to $(0, 0, 0|0, 0, 0|2, 2, 0)$.

3.2 Models

We will now turn to modeling the summary collection of joint segregating sites as described in section 3.1. We will discuss three primary models. The first model (see section 3.2.1) of interest will be used for inferring the how long ago two extant species emerged in the population. That is, we will attempt to infer the time of speciation. This model will reflect the most basic set of parameters in a two population coalescent framework and will serve as core example for most of this thesis.

The remaining two models we will consider will be extensions of the simple model. The second model (see section 3.2.7) we will be adapted to the case in which a single gene is present in the two respective populations which prevent sustained breeding between the species groups. The third and final model we will discuss (see section 3.2.12) will be adapted for the modeling of introgression between species.

3.2.1 Simple Isolation Model

This model will serve as a starting model which will commonly be the model of reference with possibly minor modifications. The simple isolation model is dependent on the following parameters

- λ : The rate of speciation between species 1 and 2.
- N_0 : The ancestral effective population size.
- N_1 : The effective population size for species 1.
- N_2 : The effective population size for species 2.

Figure 3.7 depicts the divergence of extant groups 1 and 2 from ancestral group 0 τ time units in the past. Group i ($i = 0, 1, 2$) comprises an effective number of N_i genes. While we don't explicitly model the divergence time τ in the model, we model the rate at which divergence occurs through the speciation rate parameter λ. Under the strict isolation model, the state space can be simplified further

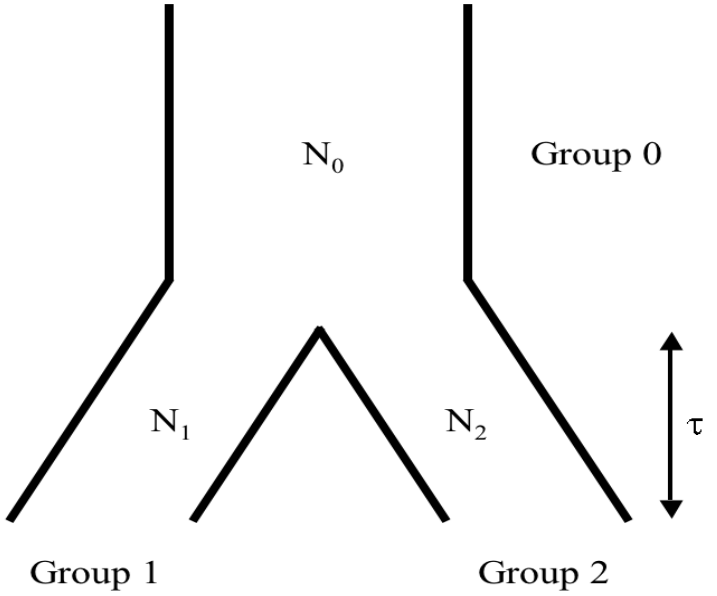

Figure 3.7: Divergence of extant groups 1 and 2 from ancestral group 0. Divergence occurs τ time units backwards in time.

since there is no possibility of obtaining states in which $(l_{1,2}, l_{1,3})$ and $(l_{2,1}, l_{2,3})$ are positive. That is, while the state is in the post-speciation phase (before the divergence event in a backwards perspective) species 1 or species 2 lineages can only have descendants in group 1 or group 2, respectively. Hence for the strict isolation model, the state space can be written as

$$\boldsymbol{S}_l = (l_{1,1}|l_{2,2}|l_{0,1}, l_{0,2}, l_{0,3}). \tag{3.1}$$

Although, the parameters (λ, N_0, N_1, N_2) are the desired parameters, it is through the mutation rate u that we are able to calibrate the events to time. Hence all the parameters are only estimable with respect to mutation rate. We consider the infinitesimal rates of change for the parameters with respect to mutation rate as

$$\Lambda = \lim \lambda/u \tag{3.2}$$

$$\theta_i = \lim 2N_i u \qquad \text{for } i \in (0, 1, 2), \tag{3.3}$$

where the limit is taken as $(\lambda, \frac{1}{N_i}, u \to 0)$. The factor of 2 in the definition of θ_i appears strictly for historical purposes. The only within level transition in this model is a speciation event, which occurs only once in the genealogy, and therefore a maximum of once per level. It is in this case that the within level transition matrix $\boldsymbol{P}_l$ is nilpotent, and the relationship given by equation (A.9) holds. Coalescence events happen at rates proportional to $1/N_i$, since the probability that two genes chosen at random share a common ancestor in the previous generation is $1/N_i$. On level l this rate is scaled by the number of ways two randomly chosen branches can be selected. That is, the total rate of coalescence on the level l is

$$\frac{\binom{l}{2}}{N_i u} = \frac{l(l-1)}{\theta_i}.$$

The rate at which mutations occur at each level l is just the rate at which a mutation can occur on a single branch u multiplied by the number of branches.

3.2.2 Coalescence Trees And Probabilities

We consider a the set of events: coalescence, speciation and mutations which explain the observed set of segregating sites observed in from a set of sampled sampled individuals. Let us explain the simple isolation model with the context of an example. Consider for a moment a data set with 3 individuals sample from group 1 and 4 individuals sampled from group 2. We let the joint segregating counts from the sampled individuals be as in table3.3 A possible evolutionary

Count	Group 1	Group 2
2	segregating	absent
3	fixed	absent
3	absent	segregating
1	absent	fixed
0	segregating	segregating
0	fixed	segregating
0	segregating	fixed

Table 3.3: Example counts of joint segregating sites.

tree explaining the mutational summary from the example data set in table 3.3 is illustrated by figure 3.8. This shows a history of 14 events. Of these 14 events, we have 9 mutation events, 4 coalescence events and the single speciation event. Under the model parameters, $\boldsymbol{\theta} = \{\lambda/u, , uN_0, uN_1, uN_2\}$, we describe the probability of the tree and the data. That is, we determine the probability $P(D, T|\boldsymbol{\theta})$, where D denotes the data and T denotes a particular tree.

In accordance with the backwards coalescence theory, discussed in section 2.2, we let the waiting times for all events in a coalescence tree be modeled as exponential distributions, with their respective rate parameters.

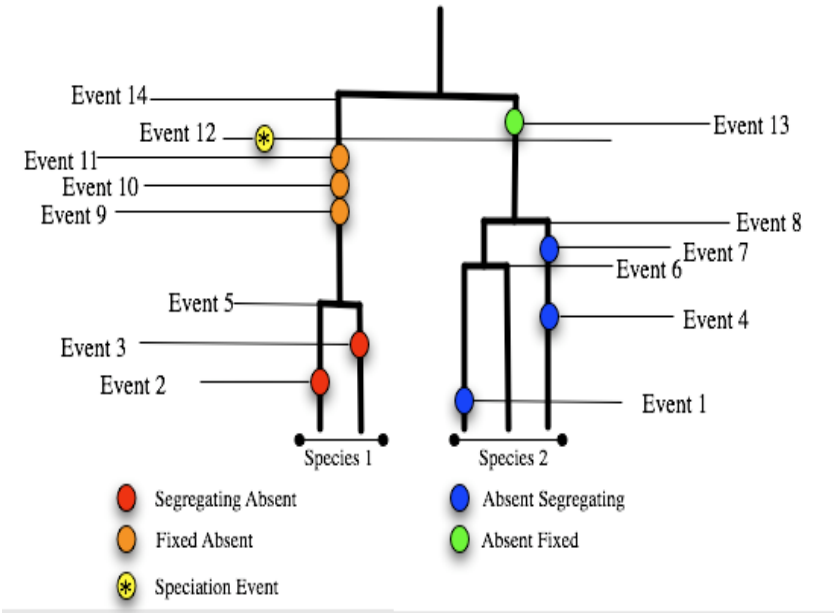

Figure 3.8: The ordered events in a genealogical history.

Since the genealogy is described in terms of an ordered history, we need the probability of this event. For the example shown in figure 3.8, the probability of the data and the genealogy is the product of the probability of each of the 14 events occurring in their illustrated orders.

Before providing the exact details on obtaining the probability of the full genealogy, we state a useful result. Letting X_{t_1} and X_{t_2} be exponentially distributed random variables with rate parameters γ_1 and γ_2 respectively. Having $t_1 \sim X_{t_1}$

and $t_2 \sim X_{t_2}$, the probability that $t_1 < t_2$ follows as

$$\begin{aligned} P(t_1 < t_2) &= \int\int_{t_1<t_2} \gamma_1 e^{-\gamma_1 t_1} \gamma_2 e^{-\gamma_2 t_2} dt_1 dt_2 \\ &= \int_0^\infty \int_0^{t_2} \gamma_1 e^{-\gamma_1 t_1} \gamma_2 e^{-\gamma_2 t_2} dt_1 dt_2 \\ &= \frac{\gamma_1}{\gamma_1 + \gamma_2}. \end{aligned}$$

The general result, when there are n competing exponentials, with rate parameters $\gamma_1, \ldots, \gamma_n$, the probability that the k^{th} event happens first is $\gamma_k / \sum_{i=1}^n \gamma_i$. We use this result to next obtain the probability of the various ordered events on the tree.

3.2.3 Transition Probabilites

Consider states of the form $(l_{1,1}|l_{2,2}|0,0,0)$, which represent a portion a genealogy in the extant post-speciation phase. The obtainable states from this initial state follows as

$$\begin{aligned} (l_{1,1}-1|l_{2,2}|0,0,0) &\qquad \text{for } l_{1,1} > 1, \\ (l_{1,1}|l_{2,2}-1|0,0,0) &\qquad \text{for } l_{2,2} > 1, \\ (0|0|l_{0,1}-1, l_{0,2}, 0) &\qquad \text{for } l_{0,1} > 1, \\ (0|0|l_{0,1}, l_{0,2}-1, 0) &\qquad \text{for } l_{0,2} > 1, \\ (0|0|l_{0,1}-1, l_{0,2}-1, 1) &\qquad \text{for } l_{0,1} > 0 \text{ and } l_{0,2} > 0, \end{aligned}$$

where we are imposing that state termination ends with a coalescence event. The last three transition events reflect those in which a speciation event occurred prior to the coalescence event. The first two transitions, representing coalescence

in groups 1 and 2 respectively, occur with probabilities

$$\frac{\frac{\binom{l_{1,1}}{2}}{N_1}}{\frac{\binom{l_{1,1}}{2}}{N_1}+\frac{\binom{l_{1,1}}{2}}{N_2}+\lambda}=\frac{\frac{l_{1,1}(l_{1,1}-1)}{\theta_1}}{\frac{l_{1,1}(l_{1,1}-1)}{\theta_1}+\frac{l_{2,2}(l_{2,2}-1)}{\theta_2}+\Lambda},$$

and

$$\frac{\frac{\binom{l_{2,2}}{2}}{N_2}}{\frac{\binom{l_{1,1}}{2}}{N_1}+\frac{\binom{l_{1,1}}{2}}{N_2}+\lambda}=\frac{\frac{l_{2,2}(l_{2,2}-1)}{\theta_2}}{\frac{l_{1,1}(l_{1,1}-1)}{\theta_1}+\frac{l_{2,2}(l_{2,2}-1)}{\theta_2}+\Lambda},$$

respectively. The last three transition events occur with probabilities

$$\frac{\lambda}{\frac{\binom{l_{1,1}}{2}}{N_1}+\frac{\binom{l_{1,1}}{2}}{N_2}+\lambda}\frac{\frac{\binom{l_{0,1}}{2}}{N_0}}{\frac{\binom{l}{2}}{N_0}}=\frac{\Lambda}{\frac{l_{1,1}(l_{1,1}-1)}{\theta_1}+\frac{l_{2,2}(l_{2,2}-1)}{\theta_2}+\Lambda}\frac{\binom{l_{0,1}}{2}}{\binom{l}{2}},$$

and

$$\frac{\lambda}{\frac{\binom{l_{1,1}}{2}}{N_1}+\frac{\binom{l_{1,1}}{2}}{N_2}+\lambda}\frac{\frac{\binom{l_{0,2}}{2}}{N_0}}{\frac{\binom{l}{2}}{N_0}}=\frac{\Lambda}{\frac{l_{1,1}(l_{1,1}-1)}{\theta_1}+\frac{l_{2,2}(l_{2,2}-1)}{\theta_2}+\Lambda}\frac{\binom{l_{0,2}}{2}}{\binom{l}{2}},$$

and

$$\frac{\lambda}{\frac{\binom{l_{1,1}}{2}}{N_1}+\frac{\binom{l_{1,1}}{2}}{N_2}+\lambda}\frac{\frac{l_{0,1}l_{0,2}}{N_0}}{\frac{\binom{l}{2}}{N_0}}=\frac{\Lambda}{\frac{l_{1,1}(l_{1,1}-1)}{\theta_1}+\frac{l_{2,2}(l_{2,2}-1)}{\theta_2}+\Lambda}\frac{l_{0,1}l_{0,2}}{\binom{l}{2}},$$

respectively.

For transitions that occur during the pre-speciationphase, that is the state representing a lineage in the ancestral state, consider the transition moving the state

$(0|0|l_{0,1}, l_{0,2}, l_{0,3})$ to one of the following states

$$\begin{aligned}
(0|0|l_{0,1}-1, l_{02}, l_{03}) &\qquad \text{for } l_{0,1} > 1,\\
(0|0|l_{0,1}, l_{02}-1, l_{03}) &\qquad \text{for } l_{0,2} > 1,\\
(0|0|l_{0,1}, l_{02}, l_{03}-1) &\qquad \text{for } l_{0,3} > 1,\\
(0|0|l_{0,1}-1, l_{02}-1, l_{03}+1) &\qquad \text{for } l_{0,1} > 0 \text{ and } l_{0,2} > 0.
\end{aligned}$$

The first transition reflects a transition in which two $l_{0,1}$ lineages coalesce, which occurs with probability

$$\frac{\frac{\binom{l_{0,1}}{2}}{N_0}}{\frac{\binom{l}{2}}{N_0}} = \frac{\binom{l_{0,1}}{2}}{\binom{l}{2}},$$

or a transition where a $l_{0,1}$ lineage coalesces with a $l_{0,3}$ lineage, which occurs with probability

$$\frac{\frac{l_{0,1}l_{0,3}}{N_0}}{\frac{\binom{l}{2}}{N_0}} = \frac{l_{0,1}l_{0,3}}{\binom{l}{2}}.$$

Hence the total probability of transition $(0|0|l_{0,1}, l_{0,2}, l_{0,3}) \to (0|0|l_{0,1}-1, l_{0,2}, l_{0,3})$ is

$$\frac{\binom{l_{0,1}}{2} + l_{0,1}l_{0,3}}{\binom{l}{2}}.$$

Similarly, the probability of transitioning to state $(0|0|l_{0,1}, l_{02}-1, l_{03})$ is

$$\frac{\binom{l_{0,2}}{2} + l_{0,2}l_{0,3}}{\binom{l}{2}}.$$

The probability of making the transition $(0|0|l_{0,1}, l_{02}, l_{03}) \rightarrow (0|0|l_{0,1}, l_{02}, l_{03}-1)$ occurs with probability

$$\frac{\frac{\binom{l_{0,3}}{2}}{N_0}}{\frac{\binom{l}{2}}{N_0}} = \frac{\binom{l_{0,3}}{2}}{\binom{l}{2}}.$$

And finally, the the transition $(0|0|l_{0,1}, l_{02}, l_{03}) \rightarrow (0|0|l_{0,1}-1, l_{02}-1, l_{03}+1)$ occurs with probability

$$\frac{\frac{l_{0,1}l_{0,2}}{N_0}}{\frac{\binom{l}{2}}{N_0}} = \frac{l_{0,1}l_{0,2}}{\binom{l}{2}}.$$

With respect to the exact recursion in probabilities, which computes the exact likelihood of the of the data (see appendix A), It is these probabilities that form the entries of the matrices $\boldsymbol{U}_l$ and $\boldsymbol{V}_l$ in the equation (A.8). $\boldsymbol{U}_l$ stores the within level probabilities $\frac{\Lambda}{\frac{l_{1,1}(l_{1,1}-1)}{\theta_1}+\frac{l_{2,2}(l_{2,2}-1)}{\theta_2}+\Lambda}$, which relates to the speciation event, and $\boldsymbol{V}_l$ stores the coalescence probabilities.

We have documented all of the transition probabilities that relate to changes in a tree's structure. In the next section, we determine the probability of the mutation configuration on a given tree.

3.2.4 Mutation Process Probabilities

In accordance with section 2.3.1, we let the (backwards) time until a mutation arises be modeled as exponential with rate parameter ul, where l denotes the level of the tree. For each mutation that arises on level l, we uniformly choose which branch it is situated on. Recall that each each level, branches are classified as type 1,2 or 3, with relative frequencies $l_1 = (l_{1,1} + l_{0,1})$, $l_2 = (l_{2,2} + l_{0,2})$ and $l_3 = l_{0,3}$. This is equivalent to modeling the time to a type i mutation as exponential with

rate parameter ul_i. On level l of a tree, the arrival of a mutation of a type i branch occurs with probability l_i/l.

Now, on level l, we are only concerned with the number of mutations that situate on type 1,2 and 3 lineages. Therefore, the order in which they arrive isn't of consequence. If we assume that q mutations arrive on level l, where q_1, q_2 and q_3 mutations are accumulated on a type i branches, then assignment of the (q_1, q_2, q_3) mutations on level l in the pre-speciation phase, has the multinomial probability

$$\frac{q!}{q_1!q_2!q_3!}\left(\frac{l_{0,1}}{l}\right)^{q_1}\left(\frac{l_{0,2}}{l}\right)^{q_2}\left(\frac{l_{0,3}}{l}\right)^{q_3}.$$

In the post-speciation phase, type 3 lineages are eliminated and the configuration of (q_1, q_2) mutations have the binomial distribution

$$\binom{q}{q_1}\left(\frac{l_{1,1}}{l}\right)^{q_1}\left(\frac{l_{2,2}}{l}\right)^{q_2},$$

for which q_1 and $q_2$mutations accumulate on type 1 and type 2 branches repsectively.

3.2.5 On Level Joint Probability Distributions

For a given level, the joint probability distribution of observing a particular arrangement of mutations and a transition to end the level, the joint probability distribution can be determined from the marginal distributions of both events. While, the mutational events are not independent of the tree since the branch types at a given level determine the types of mutations which can be distributed, they are conditionally independent, given the current state (determining tree structure) and model parameters. For instance, consider the l lineages comprising

the pre-speciation state $(0|0|l_{0,1}, l_{0,2}, l_{0,3})$, where $l = \sum_{i=1}^{3} l_{0,i}$. Accumulation of q mutations distributed as (q_1, q_2, q_3) before terminating the level with a transition to level $(0|0|l_{0,1}, l_{0,2} - 1, l_{0,3})$ occurs with probability

$$\frac{\frac{q!}{q_1!q_2!q_3!}(ul_{0,1})^{q_1}(ul_{0,2})^{q_2}(ul_{0,3})^{q_3}[\binom{l_{0,2}}{2} + l_{0,2}l_{0,3}]/N_0}{[\binom{l}{2}/N_0 + ul]^{q+1}}. \tag{3.4}$$

This probability is derived by noting that the probability of a mutation event is

$$\frac{ul}{\binom{l}{2}/N_0 + ul}.$$

Given a total of q mutations, we obtain through a product of independent probabilities

$$\left(\frac{ul}{\binom{l}{2}/N_0 + ul}\right)^{q}. \tag{3.5}$$

However, the probability of the mutational assignment (q_1, q_2, q_3) on the three branch types given the total number of mutations q has the multinomial probability

$$\frac{q!}{q_1!q_2!q_3!}\left(\frac{l_{01}}{l}\right)^{q_1}\left(\frac{l_{02}}{l}\right)^{q_2}\left(\frac{l_{03}}{l}\right)^{q_3}. \tag{3.6}$$

Hence by taking the product of (3.5) and (3.6), we obtain

$$\frac{\frac{q!}{q_1!q_2!q_3!}(ul_{0,1})^{q_1}(ul_{0,2})^{q_2}(ul_{0,3})^{q_3}}{[\binom{l}{2}/N_0 + ul]^{q}}. \tag{3.7}$$

By multiplying (3.7) by the transition probability

$$\frac{[\binom{l_{0,2}}{2} + l_{0,2}l_{0,3}]/N_0}{\binom{l}{2}/N_0 + ul},$$

we obtain the joint probability given by (3.4).

Similarly, given the post-speciation state $(l_{1,1}, l_{2,2}|0,0,0)$, accumulation of q mutations, in the arrangement (q_1, q_2) and transitioning to $(l_{1,1}, l_{2,2}-1|0,0,0)$, occurs with probability

$$\frac{\binom{q}{q_1}(ul_{0,1})^{q_1}(ul_{0,2})^{q_2}\binom{l_{2,2}}{2}/N_2}{[\lambda+\binom{l_{1,1}}{2}/N_1+\binom{l_{2,2}}{2}/N_2+ul]^{q+1}}. \tag{3.8}$$

The post-speciation state $(l_{1,1}, l_{2,2}|0,0,0)$ can also go through the speciation barrier before a coalescence event occurs. Hence, a transition accumulating q mutations in the configuration (q_1, q_2) and terminating in the state $(0|0|l_{0,1}-1, l_{0,2}-1, 1)$, occurs with probability

$$\sum_{k=0}^{q}\frac{\binom{q}{q_1}(ul_{0,1})^{q_1}(ul_{0,2})^{q_2}\lambda\frac{l_{1,1}l_{2,2}}{N_0}}{[\lambda+\binom{l_{1,1}}{2}/N_1+\binom{l_{2,2}}{2}/N_2+ul]^{k+1}[\binom{l}{2}/N_0+ul]^{q-k+1}}. \tag{3.9}$$

Equation (3.9) follows by considering all configurations of k mutations before the speciation event and the remaining $q-k$ mutations occurring in the pre-speciation phase. Since these are all disjoint events, the total probability of the q_1 and q_2 mutations arising between the boundary states, $(l_{1,1}, l_{2,2}|0,0,0)$ and $(0|0|l_{0,1}-1, l_{0,2}-1, 1)$, result from the sum of each individual probabilities with k and $q-k$ mutations before and after speciation, respectively.

3.2.6 Exact Determination Of Likelihood

Although, we have derived the joint transition probabilities on level l directly from the transition probabilities, an analysis of the recursion in pgfs $\frac{\boldsymbol{g}_l^{(\boldsymbol{p})}(\mathbf{0})}{\prod_{i=1}^{7}p_i!} = \sum_{\boldsymbol{q}}\frac{\boldsymbol{R}_l^{(\boldsymbol{q})}(\mathbf{0})}{\prod_{i=1}^{7}q_i!}\frac{\boldsymbol{g}_{l-1}^{(\boldsymbol{p}-\boldsymbol{q})}(\mathbf{0})}{\prod_{i=1}^{7}(p_i-q_i)!}$ (see appendix A) shows that the terms (3.4), (3.8), (3.9) all

result from elements in the term $\frac{\boldsymbol{R}_l^{(q)}(\mathbf{0})}{\prod_{i=1}^{7} q_i!}$. However, for exact computation of the sample, the recursion in pgfs is useful, since it gives us a format in which iterating over all tree topologies and mutational configurations is simple.

The likelihood of the data simply reflects a sum of probabilities, each representing a unique history in which the sample was derived. Also, each probability component is a product of probabilities over each level of the tree. Unfortunately, exact calculation of the probability of our sample is not easily accomplished do to the tremendous number of fully resolved trees that can explain the data. However, not all trees reflect the same degree of strength for modeling the data. In subsequent chapters, we will explore sampling techniques for constructing both the likelihood of the data and posterior distributions for the parameters in the model to help us overcome the enormous complexity involved in even the simplest of data sets.

3.2.7 Gene Incompatibility Models

While the simple isolation model (section 3.2.1) serves as a natural starting point to inferring the time since speciation, it assumes that only point mutations (base substitutions) are responsible for the genomic variation observed within a sample of individuals. However, this might not be the case.

We might consider the scenario of locus specific incompatibility factors within the genome. For instance, chromosomal rearrangements, more complicated than point substitutions, may exist and effect the locus being modeled. A paracentric inversion is a type of chromosomal rearrangement that does not include the centromere (the midsection of a chromosome) and actually permutes consecutive blocks of the genome. If a recombination event occurs during meiosis within this inverted section, the effects on the resulting chromosomes will be detrimental (see figure 3.9).

In figure 3.9, the knots on the chromosomes represent the relative position of the centromeres. In state A, we see two pairs of chromosomes which have difference genetic patterns. In state B, the child inherits a chromosome from each parent and pairs the chromosomes up in the "loop" configuration seen. State C represents the chromosomes produced after meiosis, where a recombination event occurs.

The final configuration of chromosomes is severely abnormal compared to that of the parent groups when a paracentric inversion exists. That is, two of the chromosomes that are possibly produced in meiosis phase, of the individuals with the chromosomal inversion, have either zero or two centromeres. These chromosomes with an imbalance of centromeres are a result of crossover (recombination) during the meiosis phase. This phenomena can severely cripple the offspring population,

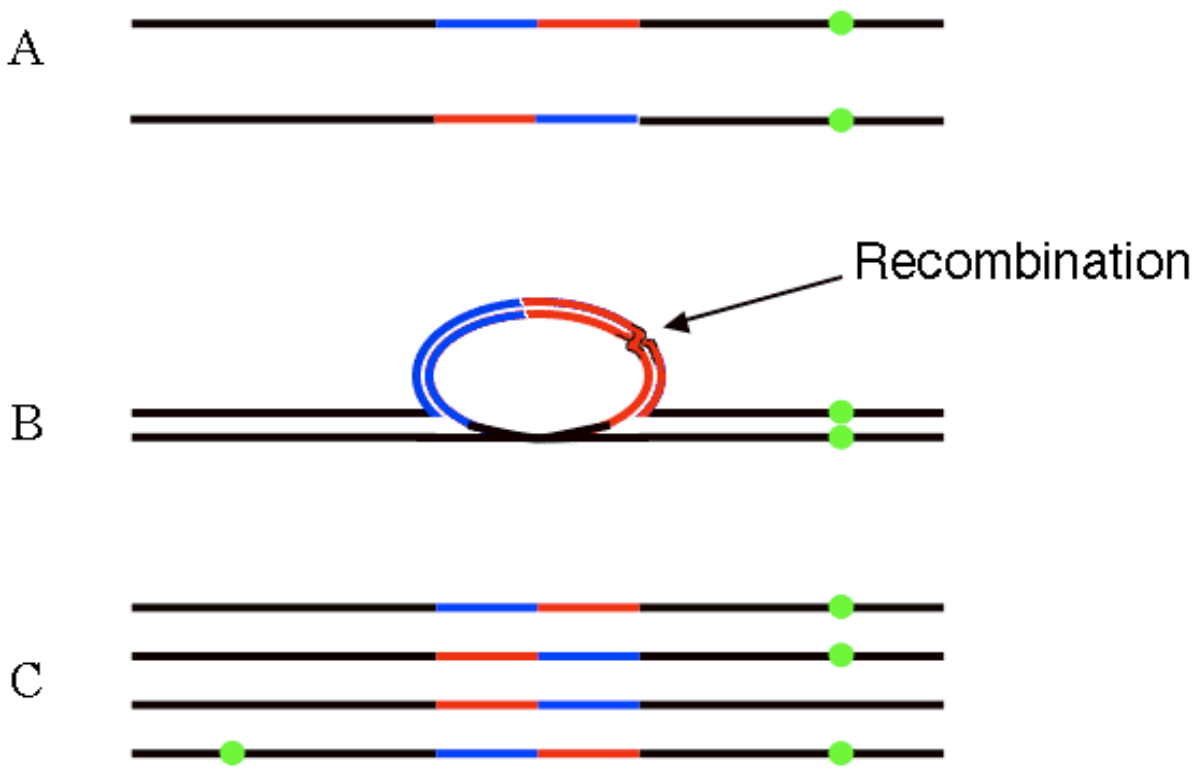

Figure 3.9: The paracentric inversion process during meiosis.

and if the recombination rate is high, lead to an isolation of the two groups. It is speculated that these chromosomal inversions, in certain species, can eventual lead to the speciation event (Noor *et al.*, 2001). However, the formation of the inversion may not coincide exactly with the emergence of a new species group, we must consider separately from the divergence.

The model we will consider is nearly identical to the simple isolation model. However, after the speciation event, the population will be separated due to an inversion existing in the two groups, which effectively prevents breeding between inverted groups. For our purposes, we will consider that group 1 descends from

the group inheriting the inversion, where the inversion is present with frquency $p \in (0, 1)$ in the ancestral population, see figure 3.10. In a backwards coalescence

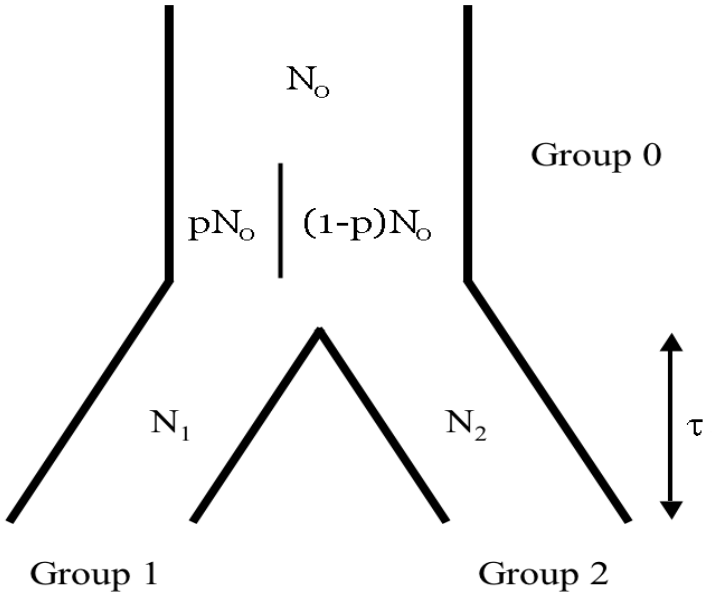

Figure 3.10: Divergence of extant groups 1 and 2 from ancestral group 0. Divergence occurs τ time units backwards in time. After the speciation barrier, there is an additional separation of groups in the ancestral state due to gene incompatibility.

prospective, the ancestral population is separated until there is only a single type 1 descendent left. That is the ancestral lineages are separated with frequency p until the state $(0|0|1, l_{0,2}, 0)$ is achieved. This states that since an inversion is fixed and absent in the respective population groups, there must be a fixed absent branch in their genealogy. Hence, even if there are no f/a (fixed/absent) mutations in the data, we condition on the presence of a f/a branch in the genealogy.

3.2.8 Transition Probabilities

While the mutational process is unperturbed by the necessity of observing an f/a branch in the history of the sample, the state transition probabilities are af-

fected. The unconditional transition probability in the extant state $(l_{1,1}|l_{22}|0,0,0)$ to $(l_{1,1}-1|l_{2,2}|0,0,0)$ is

$$\frac{\binom{l_{1,1}}{2}/N_1}{\binom{l_{1,1}}{2}/N_1+\binom{l_{2,2}}{2}/N_2+\lambda}.$$

The conditional probability at this level on seeing a fixed/absent branch in the history is augmented. For now, let us denote the conditional probability of transition and obtaining the fixed/absent branch as

$$T_e(l_{1,1}-1,l_{2,2})\frac{\binom{l_{1,1}}{2}/N_1}{\binom{l_{1,1}}{2}/N_1+\binom{l_{2,2}}{2}/N_2+\lambda}, \tag{3.10}$$

where $T_e(l_{1,1}-1,l_{2,2})$ is the probability of being in the extant state $(l_{1,1}-1|l_{22}|0,0,0)$ and observing the state $(0|0|1,l_{0,2},0)$ higher up on the tree. Similarly, we denote the probability of transitioning from $(l_{1,1}|l_{22}|0,0,0)$ to $(0|0|l_{0,1}-1,l_{0,2},0)$ conditional on seeing the fixed/absent branch in the history as

$$T_a(l_{1,1},l_{2,2})\frac{\lambda}{\binom{l_{1,1}}{2}/N_1+\binom{l_{2,2}}{2}/N_2+\lambda}\times T_a(l_{1,1}-1,l_{2,2})\frac{\binom{l_{0,1}}{2}}{\binom{l}{2}}, \tag{3.11}$$

where $T_a(i,j)$ is the conditional probability of being in the ancestral state $(0,0|i,j,0)$ and obtaining the fixed/absent branch in the history.

3.2.9 Recursion In PGFs

As presented in appendix A, we can formulate a recursion in pgfs for the coalescence model with gene incompatibility constraints. We note that the mutational process is unaffected by the transition constraints, since these happen forward in time and the rate of mutation is unperturbed by any constraints in the tree. These constraints merely reflect that at some point in the history, all the descendants

in group 1 were present with a chromosomal inversion and ultimately transmit these to their offspring. Hence the recursion in pgfs (as stated in appendix A) are modified as

$$\boldsymbol{g}_l(\boldsymbol{a}) = [\boldsymbol{I} - \boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{U}_l^*]^{-1}\boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{V}_l^*\boldsymbol{g}_{l-1}(\boldsymbol{a}) \tag{3.12}$$

where $\boldsymbol{U}_l^*$ and $\boldsymbol{V}_l^*$ denote within level and between level transitions conditioned on seeing a fixed/absent branch in the genealogy. Examples of the entries found within the matrices $\boldsymbol{U}_l^*$ and $\boldsymbol{V}_l^*$ are found in (3.10) and (3.11). The pgf matrix $\boldsymbol{F}_l(\boldsymbol{a})$ controlling the mutational process, is exactly the same as in equation (A.2).

3.2.10 Constrained Probabilites

Earlier we denoted the probability of seeing a fixed/absent branch in the genealogical path as $T_e(i,j)$ and $T_a(i,j)$ while being at extant state $(i|j|0,0,0)$ and ancestral state $(0|0|i,j,0)$, respectively. Let us now derive these probabilities through a system of recursions, which was first presented by Wiuf and Donnelly (1999).

In the ancestral state $(0|0|i,j,0)$, we have the recursion

$$T_a(i,j)\binom{i+j}{2} = T_a(i-1,j)\binom{i}{2} + T_a(i,j-1)\binom{j}{2}, \tag{3.13}$$

where the boundary conditions follow as

$$T_a(1,j) = \begin{cases} 1 & \text{for } l_{0,3} = 0 \\ 0 & \text{for } l_{0,3} > 0. \end{cases}$$

This recursion reflects a weighting scheme for transitions to states of the form $(0|0|i,j,0)$. Introducing the final boundary condition ensures these weights fall with $[0,1]$, so these are exactly the probabilities of obtaining the destination state

$(0|0|1, j, 0)$, under the coalescence framework found for the simple isolation model.

Similarly, in the extant state $(i|j|0, 0, 0)$, the recursion for finding $T_a(i, j)$ follows as

$$T_e(i,j)\left(\frac{\binom{i}{2}}{N_1}+\frac{\binom{j}{2}}{N_2}+\lambda\right)=T_e(i-1,j)\frac{\binom{i}{2}}{N_1}+T_e(i,j-1)\frac{\binom{j}{2}}{N_2}+T_a(i,j)\lambda, \quad (3.14)$$

with the boundary condition

$$T_e(1, j) = 1.$$

Hence by solving the system of recursions given by (3.13) and (3.14), we will be able to compute the probabilities, $T_e(i, j)$ and $T_a(i, j)$, of transitioning and seeing the the required state $(0|0|1, l_{0,2}, 0)$. By conditioning each of the transitions in the state space by these probabilities, as in (3.10) and (3.11), we ensure the appropriate fixed/absent branch branch exists in the genealogy.

3.2.11 Model Limitations And Usages

While we have casted the use of this coalescence model in the presence of a chromosomal rearrangement, we have failed to discus the recombination process that is associated with the paracentric inversion. Recombination, being the process in which crossover between genes occurs, can have dramatic consequences on the parameters λ, N_0, N_1, and N_2, so in general this process cannot be ignored. However, genealogy being studied is not in fact of the gene that undergoes rearrangement, but rather is *tightly linked* to the gene with the paracentric inversion and no recombination exists in the studied gene, then this is an appropriate model for inferring the demographic parameters of interest. That is, if the gene being studied shares

the same genealogy as the inverted gene, but is not subject to recombination, then the recombination process can be omitted as in the current framework. In general, we may not know of genes tightly linked with the incompatibility gene, so incorporation of recombination would be needed. However, this is not easily incorporated into a simple coalescence framework since the underlying genealogical history is of a much greater complexity than that of a simple bifurcating tree.

3.2.12 Migration Models

The simple model of isolation (section 3.2.1) can be extended to allow for introgression between groups. That is, after the species divergence, the two groups may be still allowed to exchange genes. This of course raises many questions about what a species really is. Common beliefs concerning reproductive isolation being needed for species definition are not acknowledged here. We will not be so ambitious to define what a species is, but rather we will rely on scientific expertise to define these groups.

The migration model is identical to the simple isolation model with the slight modification that groups may interchange genes from population i to population j with migration rates $m_{i,j}$ (see figure 3.11). In the migration model, there is an

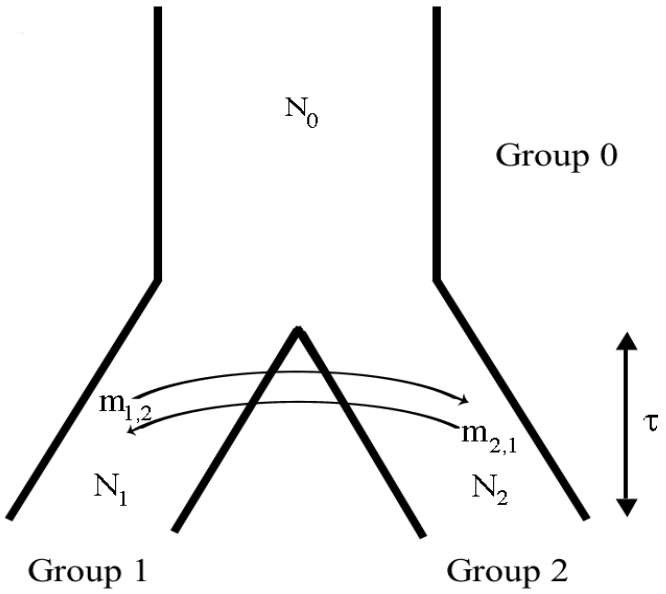

Figure 3.11: Divergence of extant groups 1 and 2 from ancestral group 0. Divergence occurs τ time units backwards in time. During the extant period, there is the possibility for gene introgression through the migration process, with migration rates $m_{1,2}$ and $m_{2,1}$.

initial population size N_0 in the ancestor species. The point of species divergence is measured as time τ, which is a backwards time measurement from the time of the present sample. At this time, the two effective population sizes N_1 and N_2 are present, and gene exchange can occur between population 1 to 2 at rate $m_{1,2}$, and from population 2 to 1 at rate $m_{2,1}$.

3.2.13 State Space And Transitions

Due to the migration process, the state space from the simple isolation model must be expanded to

$$\boldsymbol{S}_l = (l_{1,1}, l_{1,2}, l_{1,3} | l_{2,1}, l_{2,2}, l_{2,3} | l_{0,1}, l_{0,2}, l_{0,3}),$$

in order to reflect that lineages in extant populating i may have descendants of type $j \neq i$. For example, given the state $(l_{1,1}, l_{1,2}, l_{1,3} | l_{2,1}, l_{2,2}, l_{2,3} | 0, 0, 0)$, a transition can result in any of the 15 possible states. Of these, 8 of the resulting destination states reflect coalescence, 6 reflect migration, and finally one reflects speciation. Of these destination states, only the migration events are novel to our discussion this far. These possible destination states resulting from migration can be

$$\begin{aligned}
&(l_{1,1} - 1, l_{1,2}, l_{1,3} | l_{2,1} + 1, l_{2,2}, l_{2,3} | 0, 0, 0) && \text{for } l_{1,1} > 1, \\
&(l_{1,1}, l_{1,2} - 1, l_{1,3} | l_{2,1}, l_{2,2} + 1, l_{2,3} | 0, 0, 0) && \text{for } l_{1,2} > 1, \\
&(l_{1,1}, l_{1,2}, l_{1,3} - 1 | l_{2,1}, l_{2,2}, l_{2,3} + 1 | 0, 0, 0) && \text{for } l_{1,3} > 1, \\
&(l_{1,1} + 1, l_{1,2}, l_{1,3} | l_{2,1} - 1, l_{2,2}, l_{2,3} | 0, 0, 0) && \text{for } l_{2,1} > 1, \\
&(l_{1,1}, l_{1,2} + 1, l_{1,3} | l_{2,1}, l_{2,2} - 1, l_{2,3} | 0, 0, 0) && \text{for } l_{2,2} > 1, \\
&(l_{1,1}, l_{1,2}, l_{1,3} + 1 | l_{2,1}, l_{2,2}, l_{2,3} - 1 | 0, 0, 0) && \text{for } l_{2,3} > 1,
\end{aligned}$$

where the probability of transition into the first of the listed states occurs with probability

$$\frac{m_{1,2}l_{1,1}}{m_{1,2}g_{l,1} + m_{2,1}g_{l,2} + \binom{g_{l,1}}{2}/N_1 + \binom{g_{l,2}}{2}/N_2 + \lambda}, \tag{3.15}$$

where we have let $g_{l,1} = \sum_{i=1}^{3} l_{1,i}$ and $g_{2,1} = \sum_{i=1}^{3} l_{2,i}$.

3.2.14 Computational Challenges Using the Exact Method

In order to calculate the exact likelihood of the data through the recursion in pgfs

$$\boldsymbol{g}_l(\boldsymbol{a}) = [\boldsymbol{I} - \boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{U}_l]^{-1}\boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{V}_l\boldsymbol{g}_{l-1}(\boldsymbol{a})$$

as derived in equation (A.2), we must compute a sum of the form

$$\frac{\boldsymbol{g}_l^{(\boldsymbol{p})}(\boldsymbol{0})}{\prod_{i=1}^{7} p_i!} = \sum_{\boldsymbol{q}} \frac{\boldsymbol{R}_l^{(\boldsymbol{q})}(\boldsymbol{0})}{\prod_{i=1}^{7} q_i!} \frac{\boldsymbol{g}_{l-1}^{(\boldsymbol{p}-\boldsymbol{q})}(\boldsymbol{0})}{\prod_{i=1}^{7}(p_i - q_i)!}, \tag{3.16}$$

which is documented in equation (A.11). This sum reflects the marginalization of the tree space from the probability $P(D, T|\boldsymbol{\theta})$. In this recursion, the term $\boldsymbol{R}_l(\boldsymbol{a})$ has the form

$$\boldsymbol{R}_l(\boldsymbol{a}) = [\boldsymbol{I} - \boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{U}_l]^{-1}\boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{V}_l. \tag{3.17}$$

In the simple isolation model, computation of the matrix inverse $[\boldsymbol{I} - \boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{U}_l]^{-1}$ is simplified since the matrix is nilpotent. However, in the migratory case, this inverse does not simplify. the entries in the matrix $\boldsymbol{U}_l$ (appendix A) are the transition probabilities which represent migration (as in (3.15)), and

$$\frac{\lambda}{m_{1,2}g_{l,1} + m_{2,1}g_{l,2} + \binom{g_{l,1}}{2}/N_1 + \binom{g_{l,2}}{2}/N_2 + \lambda}, \tag{3.18}$$

which reflects the speciation event. The non simplification of the matrix inverse in equation (3.17) is a direct result of the fact that an infinite number of migrations

can occur before a termination of the level, by coalescence. This means that the sum used in likelihood evaluations is infinitely long and thus impossible to compute explicitly. Compare this result to the simple isolation model, where the number of possible histories, although massive, is finite. We will tackle this problem through a battery of sampling methodologies examined in subsequent chapters.

Chapter 4

Importance Sampling Approach To Parameter Estimation

We introduce an Importance Sampling (IS) procedure for estimating the population parameters ($\boldsymbol{\theta}$) which drive the evolutionary process. While these parameters may be general, we will focus on the model of speciation (see section 3.2.1). In theory, exact inference may be performed by using the recursive probability generating function

$$\frac{\boldsymbol{g}_l^{(\boldsymbol{p})}(\mathbf{0})}{\prod_{i=1}^7 p_i!} = \sum_{\boldsymbol{q}} \frac{\boldsymbol{R}_l^{(\boldsymbol{q})}(\mathbf{0})}{\prod_{i=1}^7 q_i!} \frac{\boldsymbol{g}_{l-1}^{(\boldsymbol{p}-\boldsymbol{q})}(\mathbf{0})}{\prod_{i=1}^7 (p_i - q_i)!}, \tag{4.1}$$

where $\boldsymbol{R}_l^{(\boldsymbol{q})}(\mathbf{0})$ depends on the population parameters $\boldsymbol{\theta}$. This, sum reflects a recursion in levels of a tree, over all possible mutation rearrangements on the tree. For more details, refer to section A.1.3.

While this recursion generates the exact probability of the data for a specific parameter set $\boldsymbol{\theta}$, reconstruction of the multidimensional likelihood requires too much time computing for more than one dimension. The IS sampling procedure studied in this chapter aims to speed up the computation so that a full multidi-

mensional analysis can be executed.

4.1 Importance Sampling

The goal of this chapter is to construct an importance sampler for construction of the probability of the data, where the data is the vector of mutational states (see section 3.1.3). This likelihood can be expressed by marginalizing out the tree space from the likelihood

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= P(D|\boldsymbol{\theta}) \\ &= \sum_{t\in\Omega_T} P(D,t|\boldsymbol{\theta}), \end{aligned} \tag{4.2}$$

where Ω_T is the space of all evolutionary trees which can explain the data. In terms of the exact recursion (see appendix A), this likelihood simply expresses

$$P(D|\boldsymbol{\theta}) = \frac{\boldsymbol{g}_L^{(\boldsymbol{n})}(\mathbf{0})}{\prod_{i=1}^{7} n_i!},$$

where $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5, n_6, n_7)$ is the set of mutational counts and L is the total number of individuals in the data set. If we knew how to sample from the distribution $P(D,T|\boldsymbol{\theta})$, then devising a sampling procedure for constructing the likelihood function would be simplified. That is we would generate N samples of T and D from $P(D,T|\boldsymbol{\theta})$, written as $S_t = \{t_1, t_2, \ldots, t_N\}$ and $S_d = \{d_1, d_2, \ldots, d_N\}$ respectively. The estimated likelihood function follows as

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= P(D|\boldsymbol{\theta}) \\ &\approx \frac{1}{N} \sum_{(t,d)\in(S_t,S_d)} \mathbb{I}_{(d=D)}, \end{aligned} \tag{4.3}$$

where

$$\mathbb{I}_{(d=D)} = \begin{cases} 1 & \text{if } d = D \\ 0 & \text{if } d \neq D. \end{cases}$$

Unfortunately, this method of approximating the probability of the data, under $\boldsymbol{\theta}$, is very inefficient. The inefficiency arises from the fact that hitting the observed data is a rare event which hampers the overall accuracy of the estimate when N is fixed. Using the notion of relative error, we illustrate why this is the case.

We define the relative error in our estimate (4.3) for $P(D|\boldsymbol{\theta})$ as

$$\frac{\left|\frac{1}{N}\sum_{(t,d)\in(S_t,S_d)} \mathbb{I}_{(d=D)} - P(D|\boldsymbol{\theta})\right|}{P(D|\boldsymbol{\theta})}.$$

For fixed $\epsilon > 0$, by Chebyschev's inequality $P\left(\frac{\left|\frac{1}{N}\sum_{(t,d)\in(S_t,S_d)} \mathbb{I}_{(d=D)} - P(D|\boldsymbol{\theta})\right|}{P(D|\boldsymbol{\theta})} > \epsilon\right)$ follows as

$$\begin{aligned} P\left(\left|\frac{1}{N}\sum_{(t,d)\in(S_t,S_d)} \mathbb{I}_{(d=D)} - P(D|\boldsymbol{\theta})\right| > P(D|\boldsymbol{\theta})\epsilon\right) &\leq \frac{P(D|\boldsymbol{\theta})(1-P(D|\boldsymbol{\theta}))}{NP(D|\boldsymbol{\theta})^2\epsilon^2} \\ &= \frac{1-P(D|\boldsymbol{\theta})}{NP(D|\boldsymbol{\theta})\epsilon^2}. \end{aligned}$$

This statement says, that for fixed N, the relative error will be large when $P(D|\boldsymbol{\theta})$ is small, as is the case here. Hence, the sampling method in (4.3) is expected to perform poorly. Devising a sampling method with high accuracy usually requires sampling from $P(D, T|\boldsymbol{\theta})$, such that every sample draw is consistent with the observed data.

If we knew how to sample from the distribution $P(t|D,\boldsymbol{\theta})$, then estimation of $P(\boldsymbol{\theta}|D)$ would be simple. If we sample $t_i \sim P(t|D,\boldsymbol{\theta})$ for $i = (1,2,\ldots,N)$, then

$$\begin{aligned}\sum_{t\in\Omega_T} P(D,t|\boldsymbol{\theta}) &= \sum_{t\in\Omega_T} \frac{P(D,t|\boldsymbol{\theta})}{P(t|D,\boldsymbol{\theta})} P(t|D,\boldsymbol{\theta}) \\ &\approx \frac{1}{N}\sum_{t\in S_t} \frac{P(D,t|\boldsymbol{\theta})}{P(t|D,\boldsymbol{\theta})}, \qquad (4.4)\end{aligned}$$

$$(4.5)$$

where $S_t = \{t_1, t_2, \ldots, t_N\}$. The final expression in equation (4.5) converges strongly to

$$\begin{aligned}E\left[\frac{P(D,t|\boldsymbol{\theta})}{P(t|D,\boldsymbol{\theta})}\right] &= E\left[\frac{P(D,t|\boldsymbol{\theta})P(D|\boldsymbol{\theta})}{P(t|D,\boldsymbol{\theta})P(D|\boldsymbol{\theta})}\right] \\ &= E\left[\frac{P(D,t|\boldsymbol{\theta})P(D|\boldsymbol{\theta})}{P(D,t|\boldsymbol{\theta})}\right] \\ &= E[P(D|\boldsymbol{\theta})] \\ &= P(D|\boldsymbol{\theta}).\end{aligned}$$

That is, $\forall\epsilon > 0$

$$P\left(\lim_{N\to\infty}\left|\frac{1}{N}\sum_{t\in S_T}\frac{P(D,t|\boldsymbol{\theta})}{P(t|D,\boldsymbol{\theta})} - P(D|\boldsymbol{\theta})\right| < \epsilon\right) = 1,$$

So the sum $\frac{1}{N}\sum_{t\in S_T}\frac{P(D,t|\boldsymbol{\theta})}{P(t|D,\boldsymbol{\theta})}$ closely approximates $P(D|\boldsymbol{\theta})$, for large N. Unfortunately, we do not know how to directly sample from $P(t|D,\boldsymbol{\theta})$. However, this provides us with a theoretical framework which we can exploit in constructing an efficient sampling method.

If we let $Q(D, T|\boldsymbol{\theta})$ denote a probability distribution that has the same support as $P(D, T|\boldsymbol{\theta})$, we can rewrite the probability of the data (equation 4.2) as

$$\begin{aligned} P(D|\boldsymbol{\theta}) &= \sum_{t \in \Omega_T} P(D, T|\boldsymbol{\theta}) \\ &= \sum_{t \in \Omega_T} \frac{P(D, t|\boldsymbol{\theta})}{Q(D, t|\boldsymbol{\theta})} Q(D, t|\boldsymbol{\theta}). \end{aligned} \tag{4.6}$$

In importance sampling, $Q(D, T|\boldsymbol{\theta})$ represents some distribution which we know how to sample from and is referred to as the proposal distribution (or proposal mechanism).

The data D is the classification of segregating sites vector $\boldsymbol{n}$ which contains information on both the types of segregating sites and the number of counts for each type of segregating site in the sample. We partition the data into the set

$$D = \{D_1, D_2\},$$

where D_1 represents the types of segregating site and D_2 is the counts. For example, in chapter 3, figure 3.2, the data is described by the 4 types $\{f/a, a/f, s/a, a/s\}$, with the respective counts $\{1, 1, 2, 1\}$.

We can partition the evolutionary history (T) into two components: mutations and transitions. That is we can decompose the tree structure as

$$T = \{G, U\},$$

where G represents the trees topological structure, and U represents the mutations on the tree.

Since mutations are classified by their descent, each branch in G determines which kind of mutations it can hold. Hence, knowing the types of segregating sites, given by D_1, is all we need to know when building a tree structure that is compatible with the data set.

Given a tree topology G, the mutations U can be applied which will yield a complete tree T. Therefore, the topology is sufficient for determining the branch types D_1, and the mutations and branch types are sufficient for the determination of count data D_2. Using this partition of tree space into topology and mutations, we will break our proposal mechanism $Q(\cdot)$ into two distributions $Q_1(\cdot)$ and $Q_2(\cdot)$ which pertain to proposals for the topology and mutations, respectively.

We can rewrite equation (4.6) in terms of the partitioning structure on the data and tree space, as

$$\begin{aligned} P(D|\boldsymbol{\theta}) &= \sum_{t\in\Omega_T} \frac{P(D,t|\boldsymbol{\theta})}{Q(D,t|\boldsymbol{\theta})} Q(D,t|\boldsymbol{\theta}) \\ &= \sum_{t\in\Omega_T} \frac{P(D,t|\boldsymbol{\theta})}{Q(D,t|\boldsymbol{\theta})} Q_2(D_2,U|D_1,G,\boldsymbol{\theta}) Q_1(D_1,G|\boldsymbol{\theta}), \end{aligned} \tag{4.7}$$

where

$$Q(D,T|\boldsymbol{\theta}) = Q_2(D_2,U|D_1,G,\boldsymbol{\theta}) Q_1(D_1,G|\boldsymbol{\theta}).$$

$Q_1(D_1,G|\boldsymbol{\theta})$ is the proposal mechanism for the topology G, with branch types D_1. $Q_2(D_2,U|D_1,G,\boldsymbol{\theta})$ is a proposal distribution which determines the probability of a set of mutations (U), with counts given by D_2, given a tree topology G.

The proposal mechanism $Q_1(D_1, G|\boldsymbol{\theta})$ generates a sequence

$$G = \{\boldsymbol{S}_L, \boldsymbol{S}_{L-1}, \ldots, \boldsymbol{S}_1\}$$

where $\boldsymbol{S}_i$, for $(i = L, \ldots, 1)$, represents the state of branches in the topology on level i (see section 3.1.4 for full details on the state space). We typically generate this sequence starting at S_L (the bottom of the tree) and move through the sequence

$$\boldsymbol{S}_L \rightarrow \boldsymbol{S}_{L-1} \rightarrow \cdots \rightarrow \boldsymbol{S}_1,$$

where S_1 represents the MRCA of the sample. The transition probabilities, $P(\boldsymbol{S}_{i-1}|\boldsymbol{S}_i, \boldsymbol{\theta})$ from $\boldsymbol{S}_i$ to $\boldsymbol{S}_{i-1}$, are determined by the coalescence process (see 3.2.3 for examples). The total probability for $Q_1(D_1, G|\boldsymbol{\theta})$ is determined by

$$Q_1(D_1, G|\boldsymbol{\theta}) = \prod_{i=L}^{2} P(\boldsymbol{S}_{i-1}|\boldsymbol{S}_i, \boldsymbol{\theta}).$$

Once the topology is constructed, we can superimpose the mutations, onto G,under some proposal distribution $Q_2(D_2, U|D_1, G, \boldsymbol{\theta})$.

We can approximate $P(D|\boldsymbol{\theta})$ by the sum

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= P(D|\boldsymbol{\theta}) \\ &\approx \frac{1}{N} \sum_{(d,t)\in\{S_d, S_t\}} \frac{P(d,t|\boldsymbol{\theta})}{Q(d,t|\boldsymbol{\theta})}, \end{aligned} \tag{4.8}$$

where $S_t = \{t_1, t_2, \ldots, t_N\}$ and $S_d = \{d_1, d_2, \ldots, d_N\}$ and

$$\begin{aligned} (d_i, t_i) &\sim Q(D, T|\boldsymbol{\theta}) \\ &= Q_2(D_2, U|D_1, G, \boldsymbol{\theta}) Q_1(D_1, G|\boldsymbol{\theta}). \end{aligned}$$

Even though an importance sampling scheme using this approach is feasible, it can be inefficient due to the fact that $P(D, t|\boldsymbol{\theta}) = 0$, if the data generated under $Q(D, t|\boldsymbol{\theta})$ does not conform to the observed data. Our previous partitioning of the trees into to the space $T = \{G, U\}$ is useful for generating the consistent trees, since only a tree's topology can be responsible for the data not conforming to the tree. Therefore, we must generate topologies that are consistent with D_1.

We aim to generate tree topologies under a proposal mechanism $Q_1(\cdot)$, such that

$$\begin{aligned} Q_1(G|D, \boldsymbol{\theta}) &= Q_1(G|D_1, D_2, \boldsymbol{\theta}) \\ &= Q_1(G|D_1, \boldsymbol{\theta}). \end{aligned} \tag{4.9}$$

Once a consistent tree topology is proposed, we can apply the mutations to that tree under the proposal mechanism

$$\begin{aligned} Q_2(U|D, G, \boldsymbol{\theta}) &= Q_2(U|D_1, D_2, G, \boldsymbol{\theta}) \\ &= Q_2(U|D_2, G, \boldsymbol{\theta}). \end{aligned}$$

The last line conveys that G contains all the type information in D_1. Together, $Q_1(\cdot)$ and $Q_2(\cdot)$ form the joint proposal

$$Q(T|D, \boldsymbol{\theta}) = Q_2(U|D_2, G, \boldsymbol{\theta}) Q_1(G|D_1, \boldsymbol{\theta}) \tag{4.10}$$

which enforces that $P(D, T|\boldsymbol{\theta}) > 0$. The proposal mechanism $Q_1(G|D_1, \boldsymbol{\theta})$ will generate a sequence of states

$$G = \{\boldsymbol{S}_L, \boldsymbol{S}_{L-1}, \ldots, \boldsymbol{S}_1\},$$

where the sequence is guaranteed to contain the lineage types specified by D_1. The transition probabilities $P(\boldsymbol{S}_{i-1}|\boldsymbol{S}_i, \boldsymbol{\theta})$, from the unconstrained coalescence

process, are no longer sufficient to guarantee the genealogy's consistency. We set out to constrain the process such that the observed branch set (given by D_1) is obtained in the sequence of states. We can write the joint probability (constrained such that D_1 matches the type set in the data)

$$Q_1(D_1, G|\boldsymbol{\theta}) = \prod_{i=L}^{2} T_{\mathcal{P}(\boldsymbol{S}_L)}(L_1(\boldsymbol{S}_i), L_2(\boldsymbol{S}_i), L_3(\boldsymbol{S}_i)) \times P(\boldsymbol{S}_{i-1}|\boldsymbol{S}_i, \boldsymbol{\theta}), \tag{4.11}$$

where

$$T_{\mathcal{P}(\boldsymbol{S}_L)}(L_1(\boldsymbol{S}_i), L_2(\boldsymbol{S}_i), L_3(\boldsymbol{S}_i))$$

is the probability, under the coalescent, of obtaining a topology that is consistent with the data, such that the state currently resides in the $\mathcal{P}(\boldsymbol{S}_L) \in \{\text{ancestral, extant}\} = \{a, e\}$ phase, with $(L_1(\boldsymbol{S}_i), L_2(\boldsymbol{S}_i), L_3(\boldsymbol{S}_i)$ lineages of type 1,2 and 3, respectively. In the next section, we will discuss the derivation of these probabilities.

Once we have $Q_1(D_1, G|\boldsymbol{\theta})$, we can form $Q_1(G|D_1, \boldsymbol{\theta})$ through the expression

$$\begin{aligned} Q_1(G|D_1, \boldsymbol{\theta}) &= \frac{Q_1(D_1, G|\boldsymbol{\theta})}{Q_1(D_1|\boldsymbol{\theta})} \\ &= \frac{Q_1(D_1, G|\boldsymbol{\theta})}{T_e(L_1(\boldsymbol{S}_L), L_2(\boldsymbol{S}_L), 0)}. \end{aligned} \tag{4.12}$$

Putting this altogether, the approximation to the likelihood function can be generated through the sum

$$P(D|\boldsymbol{\theta}) \approx \frac{1}{N} \sum_{t \in S_t} \frac{P(D, t|\boldsymbol{\theta})}{Q(t|D, \boldsymbol{\theta})}, \tag{4.13}$$

where $S_t = \{t_1, t_2, \ldots, t_N\}$ with $t_i \sim Q(t|D, \boldsymbol{\theta})$, and $Q(t|D, \boldsymbol{\theta})$ has the form shown in equation (4.10).

The next section discusses in detail the proposal distributions $Q_1(\cdot)$ and $Q_2(\cdot)$, for sampling consistent trees.

4.2 Proposal Mechanisms

We propose samples under a two stage partition of the genealogical history. We partition the history $T = \{G, U\}$, where G represents the topology of the history and U is the mutational history given the topology. Given $\boldsymbol{\theta}$, we can sample from G directly by simulating a coalescence path, from the sample, conditional on having the correct topological branch structure so that $P(D_1|G) > 0$ through the process shown in equation (4.11).

From the partition $T = \{G, U\}$ and the partition of the data $D = \{D_1, D_2\}$ into types and counts, we describe an importance sampling approximation to the likelihood function of the form

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= P(D|\boldsymbol{\theta}) \\ &\approx \frac{1}{N} \sum_{t \in S_t} \frac{P(D, t|\boldsymbol{\theta})}{Q(t|D, \boldsymbol{\theta})}, \end{aligned}$$

where

$$Q(T|D, \boldsymbol{\theta}) = Q_2(U|D_2, G, \boldsymbol{\theta}) Q_1(G|D_1, \boldsymbol{\theta}).$$

We set out to derive a set of model based proposal distributions $Q_1(\cdot)$ and $Q_2(\cdot)$ that determine the topological history and mutational placements of the sample respectively. While the mutational placements is the most difficult component of the sampling mechanism, the topology solely determines which trees are compatible with the mutations seen in the sample so conditioning the mutational placements on consistent topologies is a natural approach. We will first show how to sample from $Q_1(G|D_1, \boldsymbol{\theta})$ then, conditional on G, show how to sample from $Q_2(U|D_2, G, \boldsymbol{\theta})$.

4.2.1 Sampling Histories With Topological Constraints

Some topologies are inconsistent with the data since they do not contain the sufficient branch types in that have the necessary lineages.

For example, if a data set contains a fixed-absent mutation (f/a), there must be a branch present that can hold that mutation. That is, there must be a branch in the tree that has descendants in all of the species 1 group and none in the species 2 group. It is also the case that some branches in the topology restrict the presence of other branch types in the topology. For example, if a fixed-absent branch is present in the history, then a segregating-fixed branch may not be present. Table 4.1 summarizes all the type incompatibilities.

Type	Group 1/Group 2	Incompatible	Compatible
1	s/a	None	All
2	f/a	$s/s, s/f$	$\{s/a, a/s\}$ and $\{a/f$ or $f/s\}$
3	a/s	None	All
4	a/f	$s/s, f/s$	$\{s/a, a/s\}$ and $\{f/a$ or $s/f\}$
5	s/s	$f/a, a/f$	$\{s/a, a/s\}$ and $\{f/s$ or $s/f\}$
6	f/s	$a/f, s/f$	$\{s/a, a/s\}$ and $\{f/a$ or $s/s\}$
7	s/f	$f/a, f/s$	$\{s/a, a/s\}$ and $\{a/f$ or $s/s\}$

Table 4.1: Incompatibilities between types.

Given the type incompatibilities, we can partition the space of trees into four mutually exclusive structures, which are depicted in figure 4.1. The partition on

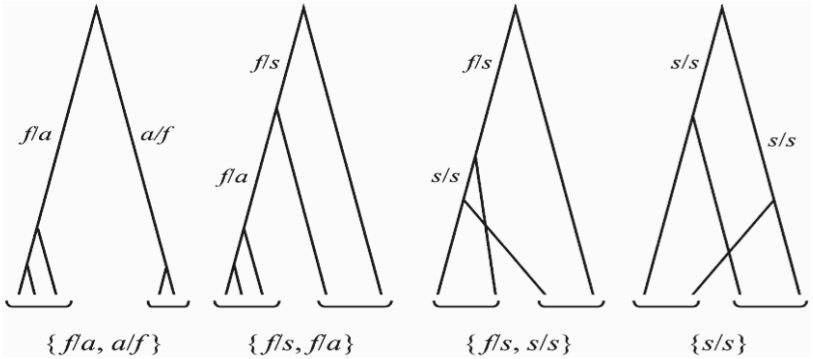

Figure 4.1: The four partitions of tree space based on branch topologies (figure from Leman *et al.* (2005)).

tree space shows that trees will either have branches in the history that are of type $\{f/a, a/f\}$, $\{f/s, f/a\}$, $\{f/s, s/s\}$ or $\{s, s\}$. We assume under this partition that the labels of the species types (1 and 2) are arbitrary and can be exchanged if necessary.

4.2.2 Example: Trees With Fixed-Absent And Absent-Fixed Lineages

After seeing the data set and observing which types are present, we can choose which type of tree is eligible for explaining the data. This is how we will propose trees which are guaranteed to be compatible for the data. In theory, we could run an Accept Reject (AR) algorithm to sample consistent trees from the coalescence process, but this will reject a large proportion of the trees. Consider for example, the space of topologies made up from two individuals in each species group. Figure 4.2 shows the space of topologies. For this relatively small tree space, we notice that not all of the trees have fixed-absent and absent-fixed lineages. In fact, only trees $1, 3, 4, 6, 10$ and 11 have lineages that would be consistent with the observation of a f/a and a/f mutation. The probabilities of each of these 6 topologies follows as

$$\begin{aligned}
P(T_1) &= \frac{\frac{1}{N_1}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} \times \frac{\frac{1}{N_2}}{\frac{1}{N_2}+\lambda} \\
P(T_3) &= \frac{\frac{1}{N_1}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} \times \frac{\lambda}{\frac{1}{N_2}+\lambda} \times \frac{1}{3} \\
P(T_4) &= \frac{\frac{1}{N_2}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} \times \frac{\frac{1}{N_1}}{\frac{1}{N_1}+\lambda} \\
P(T_{10}) &= \frac{\lambda}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} \times \frac{1}{6} \times \frac{1}{3} \\
P(T_{11}) &= \frac{\lambda}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} \times \frac{1}{6} \times \frac{1}{3}
\end{aligned}$$

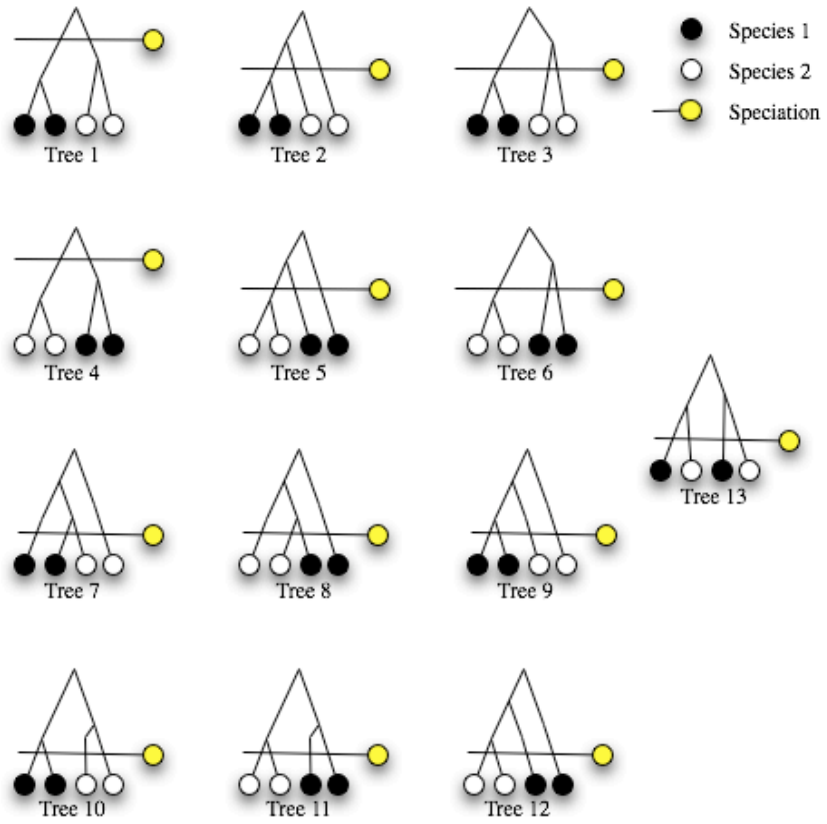

Figure 4.2: 13 topologies are possible when there are 2 individuals in each species group.

In the case where $\lambda = N_1 = N_2 = 1$, the total probability of all of these trees is 0.4815, which is the probability of obtaining a tree with fixed and absent lineages in both groups. In our earlier notation, for the weighting probabilities $T_{\mathcal{P}(\boldsymbol{S}_L)}(L_1(\boldsymbol{S}_i), L_2(\boldsymbol{S}_i), L_3(\boldsymbol{S}_i))$, we would have written

$$T_e(2, 2, 0) = 0.4815.$$

So far, we have obtained this value only through the enumeration of the entire tree space, which is infeasible for most problems. However, the next section details the derivation of these probabilities by a recursion through the tree levels.

4.2.3 Recursion Constraints

Since the coalescence process is a hierarchical process, from tree level to tree level, the entering level can be weighted by the probability of ultimately obtaining a particular branch type. Hence, if the entering level will restrict the particular branch types that are needed to make the tree compatible, the entering level will have probability zero and will not be selected. Wiuf and Donnelly (1999) derived a recursive form for weighting the coalescence process such that a f/a branch is present in the history. Leman *et al.* (2005) extend the result of Wiuf and Donnelly (1999) to condition on the existence of the branch types $\{f/a, a/f\}$, $\{f/s, f/a\}$, $\{f/s, s/s\}$ and $\{s, s\}$, so that for any (nonrecombining) data set, a consistent tree may be proposed.

We lay out the recursion probabilities for weighting the entering levels of the tree, so that each proposed tree will be consistent with the data set. Since there are four possible topological configurations on the tree space, four sets of probability weights are required.

Before we layout the recursion, let us illustrate the general procedure for the case when $\lambda = N_1 = N_2 = 1$ and compute $T_e(2, 2, 0)$.

4.2.4 Recursion Constraints Example

We refer to the example in section 4.2.2 and illustrate and efficient method for computing the probability of compatible trees based on a sequence of recursively determined calculations.

From the state $\boldsymbol{S}_4 = (2,2|0,0,0)$, the possible transitions are to states $(1,2|0,0,0)$, $(2,1|0,0,0)$ and $(0,0|2,2,0)$ which occur with probabilities $\frac{\frac{1}{N_1}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda}$, $\frac{\frac{1}{N_2}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda}$ and $\frac{\lambda}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda}$, respectively. Of course, each of these entry states induces some probability of hitting a tree with fixed and absent lineages in both groups. This determines the recursion

$$T_e(2,2,0) = T_e(1,2,0)\frac{\frac{1}{N_1}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} + T_e(2,1,0)\frac{\frac{1}{N_2}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} + T_a(2,2,0)\frac{\lambda}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda}.$$

Following this procedure for the determination of $T_e(1,2,0)$ yields

$$T_e(1,2,0) = T_e(1,1,0)\frac{\frac{1}{N_2}}{\frac{1}{N_2}+\lambda} + T_a(1,2,0)\frac{\lambda}{\frac{1}{N_2}+\lambda}.$$

Now since, $\boldsymbol{S}_2 = (1,1|0,0,0)$ has lineages which are fixed in both groups, $T_e(1,1,0) = 1$, which results in

$$T_e(1,2,0) = \frac{\frac{1}{N_2}}{\frac{1}{N_2}+\lambda} + T_a(1,2,0)\frac{\lambda}{\frac{1}{N_2}+\lambda}$$

. Similarly, we have

$$T_e(2,1,0) = \frac{\frac{1}{N_1}}{\frac{1}{N_1}+\lambda} + T_a(2,1,0)\frac{\lambda}{\frac{1}{N_2}+\lambda}.$$

All that is left is to find $T_a(2,1,0)$, $T_a(1,2,0)$ and $T_e(2,2,0)$.

If we arrived at the state $\boldsymbol{S}_4 = (0, 0|2, 2, 0)$, then a coalescence between any group is possible under the coalescence process. This results in the recursion

$$T_a(2, 2, 0) = T_a(1, 2, 0)\frac{1}{6} + T_a(2, 1, 0)\frac{1}{6} + T_a(1, 1, 1)\frac{4}{6}.$$

Note however that the state $\boldsymbol{S}_3 = (0, 0|1, 1, 1)$ violates the existence of lineages which are fixed and absent in both groups. Figure 4.3 illustrates why lineages can't be fixed and absent in both groups if a type 3 lineage exists before all type 1 and type 2 lineages coalesce amongst themselves.

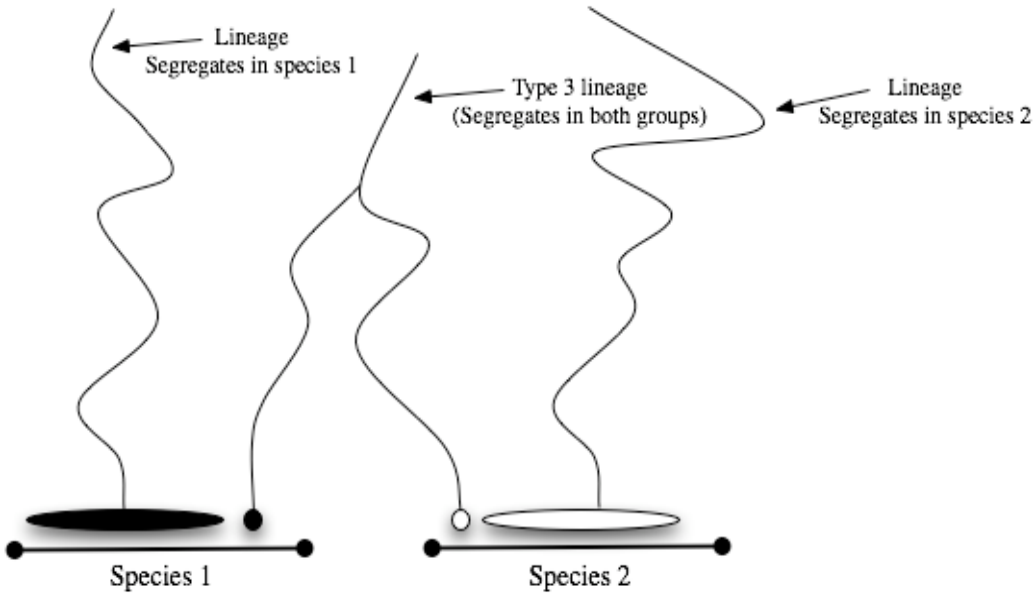

Figure 4.3: Fixed absent lineages cannot exist if type 3 lineages emerge prior to coalesce of all type 1 and type 2 lineages amongst themselves.

From this we determine that $T_a(1, 1, 1) = 0$ and hence

$$T_a(2, 2, 0) = T_a(1, 2, 0)\frac{1}{6} + T_a(2, 1, 0)\frac{1}{6}.$$

Since $T_a(1,1,0) = 1$ (i.e. fixed absent lineages exist in both groups), we have that $T_a(2,1,0) = T_a(1,2,0) = 1/3$. Plugging all of these probabilities into their appropriate spots yields the final probability that

$$T_e(2,2,0) = \frac{\frac{\frac{1}{N_2}+\lambda/3}{\frac{1}{N_2}+\lambda}\frac{1}{N_1}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} + \frac{\frac{\frac{1}{N_1}+\lambda/3}{\frac{1}{N_1}+\lambda}\frac{1}{N_2}}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda} + \frac{\lambda/9}{\frac{1}{N_1}+\frac{1}{N_2}+\lambda}.$$

Upon plugging in the values $\lambda = N_1 = N_2 = 1$, we obtain $T_e(2,2,0) = 0.4815$, which coincides with our earlier calculation.

We will now derive these recursion probabilities in the most general forms for all four of the topologies shown in figure 4.1

4.2.5 Case 1: Trees with $\{f/a, a/f\}$ Branches

In the case where we require that a topology has f/a and a/f branches, we require that the tree have fixed lineages in both groups, the MRCA of the history may only be obtained after all of the lineages of type 1 and 2 have coalesced among themselves (and not with each other). Letting $T_a(i,j,k)$ denote the probability of obtaining f/a and a/f branches in the ancestral portion of the history when there are i lineages of type 1, j lineages of type 2, and k lineages of type 3, we have that

$$T_a(i,j,k) = 0 \qquad \text{if } k > 0 \text{ and } i,j > 1,$$

which states that all type 1 and 2 lineages must coalesce prior to the formation of a type 3 lineage. Recursively, we express the probability of obtaining the $\{f/a, a/f\}$

branch set as

$$T_a(i,j,0) = T_a(i-1,j,0)\left(\frac{\binom{i}{2}}{N_0}/\frac{\binom{i+j}{2}}{N_0}\right) + T_a(i,j-1,0)\left(\frac{\binom{j}{2}}{N_0}/\frac{\binom{i+j}{2}}{N_0}\right). \quad (4.14)$$

Equation (4.14) expresses the probability, while in the ancestral phase, of obtaining the branch set $\{f/a, a/f\}$ through the possible entry states it can make while having i type 1 lineages and j type 2 lineages. Hence, either coalescence of a type 1 linage can occur (with probability $\frac{\binom{i}{2}}{N_0}/\frac{\binom{i+j}{2}}{N_0}$), or a coalescence of a type 2 lineage can occur (with probability $\frac{\binom{j}{2}}{N_0}/\frac{\binom{i+j}{2}}{N_0}$). Each of these transitions leaves the process at a state with either $(i-1,j,0)$ or $(i,j-1,0)$ type 1,2 and 3 lineage types, respectively. We can rewrite equation (4.14) as

$$T_a(i,j,0)\binom{i+j}{2} = T_a(i-1,j,0)\binom{i}{2} + T_a(i,j-1,0)\binom{j}{2}. \quad (4.15)$$

In the ancestral phase, since the obtainment of the branch types $\{f/a, a/f\}$ will be satisfied once the state has $(1,1,0)$ type 1,2 and 3 lineage types, we have that

$$T_a(1,1,0) = 1.$$

We let $T_e(x,y,z)$ denote the probability of obtaining f/a and a/f branches in the extant portion of the history when there are i lineages of type 1, j lineages of type 2, and k lineages of type 3. It follows that

$$T_e(i,j,k) = 0 \qquad \text{if } k > 0.$$

If we are dealing with a non-migrating process, it is guaranteed that $z > 0$. The recursion for $T_e(x, y, 0)$ follows as

$$T_e(i,j,0)\left(\frac{\binom{i}{2}}{N_1}+\frac{\binom{j}{2}}{N_2}+\lambda\right)=T_e(i-1,j,0)\frac{\binom{i}{2}}{N_1}+T_e(i,j-1,0)\frac{\binom{j}{2}}{N_2}+T_a(i,j,0)\lambda \tag{4.16}$$

with the boundary condition

$$T_e(1,1,0)=1.$$

4.2.6 Case 2: Trees with $\{f/a, f/s\}$ Branches

For the case where we require $\{f/a, f/s\}$ branches in the history of the sample, we derive a recursive system that results in the required weighting probabilities for the coalescence process. Since a state in which there are $(i, 1, 0)$ type 1,2 and 3 linages, results in an a/f branch, which is incompatible with the f/s mutation, we have

$$T_a(i,1,0)=T_e(i,1,0)=0 \qquad \text{for } i>0. \tag{4.17}$$

$T_a(i, j, k)$ and $T_e(i, j, k)$ denote the probabilities of obtainment of the $\{f/a, f/s\}$ branch set, at states with (i,j,k) type 1,2 and 3 lineage types, in the ancestral and extant phases of the process, respectively.

Since a f/a branch requires a state with lineage types $(1, j, 0)$ and a f/s branch requires a state, in the ancestral phase, with lineage types $(0, j, k)$, we have that

$$T_a(i,j,k)=0 \qquad \text{for } ik\neq 0. \tag{4.18}$$

Because a f/a branch must arise previous to the f/s under the coalescence process, we must obtain the a state with $(1, k, 0)$ type 1,2 and 3 lineages, respectively, prior to obtainment of a state with lineage numbers $(0, j, 1)$ for $j \geq 1$. Hence we solve the recursions

$$T_a(i, j, 0)\binom{i+j}{2} = T_a(i-1, j, 0)\binom{i}{2} + T_a(i, j-1, 0)\binom{j}{2}, \tag{4.19}$$

and

$$T_e(i, j, 0)\left(\frac{\binom{i}{2}}{N_1} + \frac{\binom{j}{2}}{N_2} + \lambda\right) = T_e(i-1, j, 0)\frac{\binom{i}{2}}{N_1} + T_e(i, j-1, 0)\frac{\binom{j}{2}}{N_2} + T_a(i, j, 0)\lambda, \tag{4.20}$$

while respecting the boundary conditions in equations (4.17) and (4.18). The boundary condition in equation (4.18) will guarantee that a f/a branch is observed prior to a f/s branch. Once the f/a branch arises, the recursion in equation (4.19) changes to

$$T_a(1, j, 0)\binom{i+j}{2} = T_a(i, j-1, 0)\binom{j}{2} + T_a(0, j-1, 1)j. \tag{4.21}$$

At the point in the process where there are $(0, j, 1)$, for $j \geq 1$, type 1,2 and 3 lineages, respectively, we have

$$T_a(0, j, 1) = 1.$$

4.2.7 Case 3: Trees with $\{f/s, s/s\}$ Branches

For the case where we require $\{f/s, s/s\}$ branches in the history of the sample, we derive a recursive system that results in the required weighting probabilities for the coalescence process. We let $T_a(i, j, k)$ be the probability, while in the ancestral group, of obtaining the $\{f/s, s/s\}$ branch set when there are (i, j, k) lineages of type 1,2 and 3. Since at least one a/s branch must exist on any level in which a f/s branch exists, we have that

$$T_a(i, 0, k) = 0 \qquad \text{for } i, j \geq 0. \tag{4.22}$$

Another way to see this is, if there are type 1 and type 3 lineages present, then in order to form a f/s branch, all the type 3 and type 1 lineages must coalesce, however a type 2 lineage must still be present for there to exist a branch with segregating descendants in the second group.

Also, we have that

$$T_e(i, j, 0) = T_a(i, j, 0) = 0 \qquad \text{for } i, j < 2, \tag{4.23}$$

where $T_e(i, j, k)$ is the probability, while in the extant group, of obtaining the $\{f/s, s/s\}$ branch set when there are (i, j, k) lineages of type 1,2 and 3. This follows since a lineage of both types 1 and 2 must still be present when a lineage of type 3 forms in order for there to exist branches that have segregating descendants in both groups.

Hence, we must solve the recursion

$$T_e(i,j,0)\left(\frac{\binom{i}{2}}{N_1}+\frac{\binom{j}{2}}{N_2}+\lambda\right)=T_e(i-1,j,0)\frac{\binom{i}{2}}{N_1}+T_e(i,j-1,0)\frac{\binom{j}{2}}{N_2}+T_a(i,j,0)\lambda, \tag{4.24}$$

while respecting the boundary condition in equations (4.22) and (4.23). The recursion for $T_a(i,j,0)$ follows as

$$T_a(i,j,0)\binom{i+j}{2}=T_a(i-1,j,0)\binom{i}{2}+T_a(i,j-1,0)\binom{j}{2}+T_a(i-1,j-1,1)ij. \tag{4.25}$$

Under the formation of a type 3 lineage (by the coalescence of a type one and type 2 lineages), the recursion for $T_a(i,j,k)$ follows as

$$\begin{aligned} T_a(i,j,k)\binom{i+j+k}{2} &= T_a(i-1,j,k)\left(\binom{i}{2}+ik\right)+T_a(i,j-1,0)\left(\binom{j}{2}+jk\right) \\ &+ T_a(i-1,j-1,k+1)ij+T_a(i,j,k-1)\binom{k}{2}. \end{aligned} \tag{4.26}$$

Hence, solving the recursive system of equations (4.24), (4.25) and (4.26) yields the probability weights for the coalescence process that ensures consistent trees with the $\{f/s, s/s\}$ branch set.

The boundary condition (4.23) will enforce that a state $T(i,j,k)$ is achieved for $i \geq 1, j \geq 1$ and $k > 0$, which produces an s/s branch. The boundary condition in equation (4.22) will produce the f/s branch . This will result in the lineages $(0,1,1)$ at which point

$$T_1(0,1,1) = 1.$$

4.2.8 Case 4: Trees with $\{s/s\}$ Branches

We now focus on the sampling weights that ensure the presence of a branch with descendants that segregate in both groups. We also require that no f/s or s/f branches can exist in this topology since this is covered by case 3. We let $T_a(i,j,k)$ be the probability, while in the ancestral phase, of obtaining a s/s branch in the topology at a state where there are (i,j,k) type 1,2 and 3 lineages, respectively. Similarly, we let $T_e(i,j,k)$ be the probability, while in the extant phase, of obtaining a s/s branch in the topology at a state where there are (i,j,k) type 1,2 and 3 lineages, respectively.

Since the presence of an s/s branch forbids the presence of f/a and a/f branches, we have the boundary conditions that

$$T_a(i,1,0) = T_e(i,1,0) = 0$$

forbids presence of a f/a branch, and

$$T_a(1,j,0) = T_e(1,j,0) = 0$$

which forbids the presence of an a/f branch. Also, since we are not allowing for branches which have fixed descendants in one group and segregate in the other, we have the boundary conditions

$$T_a(i,0,1) = T_a(0,j,1) = 0.$$

The recursion for $T_a(i,j,0)$ follow as

$$T_a(i,j,0)\binom{i+j}{2} = T_a(i-1,j,0)\binom{i}{2} + T_a(i,j-1,0)\binom{j}{2} + T_a(i-1,j-1,1)ij.$$

Under the formation of a type 3 lineage, the recursion for $T_a(i,j,k)$ follows as

$$\begin{aligned} T_a(i,j,k)\binom{i+j+k}{2} &= T_a(i-1,j,k)\left(\binom{i}{2}+ik\right)+T_a(i,j-1,0)\left(\binom{j}{2}+jk\right) \\ &+ T_a(i-1,j-1,k+1)ij+T_a(i,j,k-1)\binom{k}{2}. \end{aligned}$$

In the extant phase of the process, the recursion for $T_e(i,j,k)$ follows as

$$T_e(i,j,0)\left(\frac{\binom{i}{2}}{N_1}+\frac{\binom{j}{2}}{N_2}+\lambda\right)=T_e(i-1,j,0)\frac{\binom{i}{2}}{N_1}+T_e(i,j-1,0)\frac{\binom{j}{2}}{N_2}+T_a(i,j,0)\lambda.$$

Once a type 3 lineage is present in the history, we have

$$T_a(0,0,k)=1$$

and the standard unconstrained coalescence process is performed on the remaining lineages.

4.2.9 Sampling The Mutations

The sampling distribution $P(U|G)$ is complicated and direct sampling cannot be done under the true coalescence model. However, the number of accumulated mutations on any branch in G is expected to be larger than those on shorter branches. While we don't explicitly track the branch lengths in our topology (i.e. they have been integrated out of the model), we can indirectly discuss them. Since we are tracking the ordered coalescence and speciation events in the topology G, which occur via a structured Poisson process, the time between events is exponentially distributed. This is all described at large in chapter 3.

In the extant post-speciation phase of the process, with (l_1, l_2, l_3) type 1,2 and 3 lineages, the time to the first transition follows as

$$t = \min(t_1, t_2, t_3), \tag{4.27}$$

where

$$\begin{aligned} t_1 &\sim \exp\left(\binom{l_1}{2}/N_1\right) \\ t_2 &\sim \exp\left(\binom{l_2}{2}/N_2\right) \\ t_3 &\sim \exp\left(\lambda\right). \end{aligned}$$

Since t_1, t_2 and t_3 follow independent exponential distributions, t (shown by equation (4.27)) also follows an exponential distribution with the sum of the rate parameters. The distribution of the minimum time t from equation (4.27) follows as

$$t \sim \exp\left(\binom{l_1}{2}/N_1 + \binom{l_2}{2}/N_2 + \lambda\right). \tag{4.28}$$

In the ancestral phase of the process, the time spent on level l follows as

$$t \sim \exp\left(\binom{l}{2}/N_0\right). \tag{4.29}$$

Recall that the time until a mutation event arises also follows a an exponential distribution (see sections 3.2.4 and A.1.1). Therefore, the distribution of mutations, on any level of the distribution, is directly related to the amount of time spent in that level. Since longer branches are expected to accumulate more mutations, our proposal distribution for the placement of mutations ($Q_2(\cdot)$) is regulated by the expected length of time spent in a certain level.

The expected time spent on each level is computed as the inverse of the rate parameters from equations (4.28) and (4.29). That is, the expected length of each level in the prespeciation phase is

$$w_l = N_0 / \binom{l}{2}.$$

In the postspeciation phase, the expected length is

$$w_l = 1/\left[\frac{\binom{\sum_{i=0}^{2} l_{1,i}}{2}}{N_1} + \frac{\binom{\sum_{i=0}^{2} l_{2,i}}{2}}{N_2} + \lambda\right].$$

Consider for example the type 1 (s/a) mutations. We let $e_{l,1}$ be the number of eligible type 1 lineages, on level l of the topology G, and let U_1 denote the mutational placement of type 1 (s/a) mutations. We propose the placement of s/a mutations in the data set under the multinomial distribution

$$Q_2(U_1|D_2, G, \boldsymbol{\theta}) = n_1! \prod_{l=2}^{L} \frac{(r_{l,1})^{n_{1,l}}}{n_{1,l}!}, \tag{4.30}$$

where $n_{1,l}$ is the number of type 1 mutations placed on level l. The weighting probability $r_{l,1}$ is proportional to the expected length of the level and the number of eligible lineages $e_{l,1}$. Explicitly, we write this as

$$r_{1,l} = \frac{e_{l,1} w_l}{\sum_{k=2}^{L} e_{k,1} w_k}. \tag{4.31}$$

Let us give an explicit example of how this sprinkling procedure works. Consider

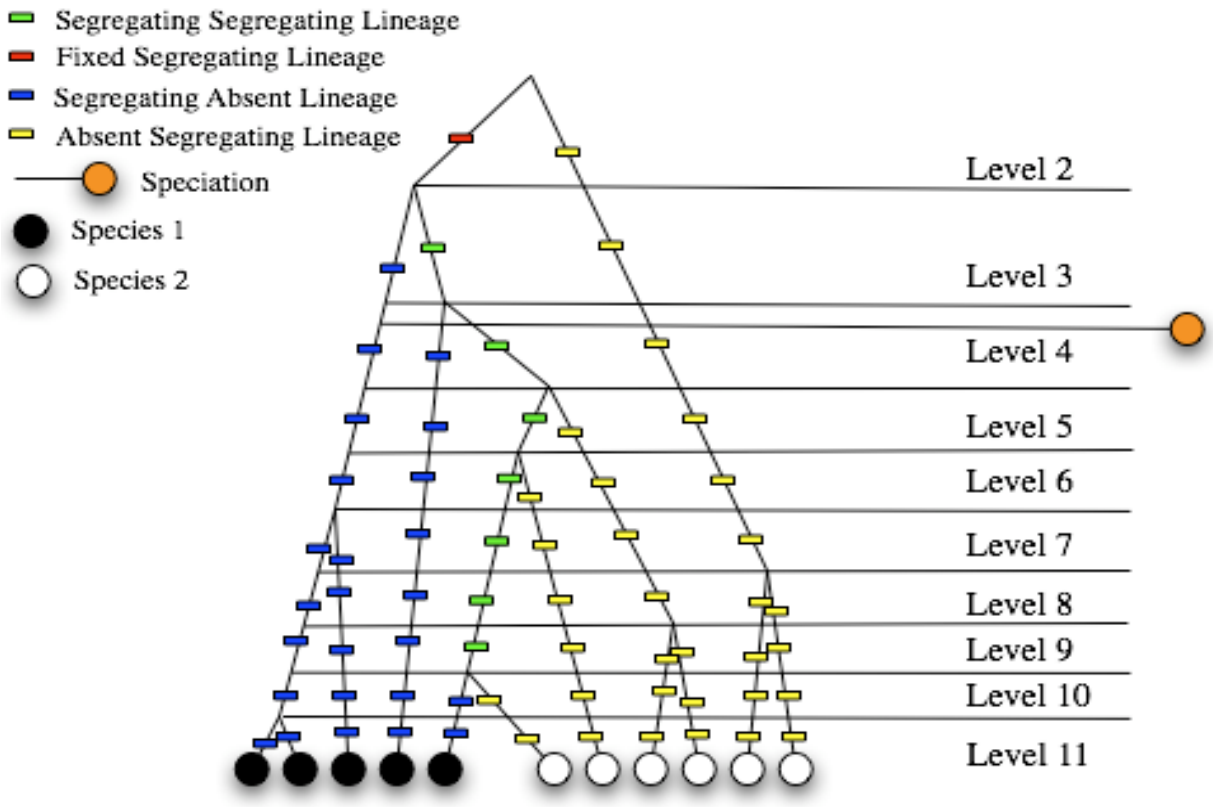

Figure 4.4: Tree structure with fixed-segregating and segregating-segregating lineages..

the tree structure in figure 4.4, where each lineage on level l has been marked for clarity.

For sprinkling the type 1 segregating-absent (s/a) mutations, we use the sampling distribution in equation (4.30), where $r_{1,l}$ is a weighting of the eligible lineages (see equation (4.31)). For example, on level 11, we have

$$e_{11,1}w_{11} = 5 \times \frac{1}{\frac{\binom{5}{2}}{N_1} + \frac{\binom{6}{2}}{N_2} + \lambda}.$$

For all 7 types of mutations, we write the full summary collection as $U = \{U_1, U_2, \ldots, U_7\}$.

From this, the full joint distribution for the placement of all mutations in D on a given G is expressed as the product of multinomial distributions. This joint multinomial distribution follows as

$$\begin{aligned} Q_2(U|D_2, G, \boldsymbol{\theta}) &= \prod_{j=1}^{7} Q_2(U_j|D_2, G, \boldsymbol{\theta}) \\ &= \prod_{j=1}^{7} n_j! \prod_{l=2}^{L} \frac{(r_{l,j})^{n_{j,l}}}{n_{j,l}!}, \end{aligned} \tag{4.32}$$

where

$$r_{j,l} = \frac{e_{l,j} w_l}{\sum_{k=2}^{L} e_{k,j} w_k}.$$

From this, we have the full joint proposal distribution

$$Q(T|D, \boldsymbol{\theta}) = Q_2(U|D_2, G, \boldsymbol{\theta}) Q_1(G|D_1, \boldsymbol{\theta}).$$

4.3 Tuning The Proposal

In our description of the proposal mechanism

$$Q(T|D,\boldsymbol{\theta}) = Q_2(U|D_2,G,\boldsymbol{\theta})Q_1(G|D_1,\boldsymbol{\theta}),$$

we are conditioning the proposal of genealogies T on a fixed set of $\boldsymbol{\theta}$. We call this fixed set of $\boldsymbol{\theta}$ the driving set of parameters, or simply *driving set* or *driving values* for the proposal. A driving set $\boldsymbol{\theta}$ that is not in the mass of the likelihood function, that is, in the concentrated region of the likelihood function, will create a poor proposal distribution. This arises since, under the observed data, an *unlikely* $\boldsymbol{\theta}$ set will produce low probability trees.

Since our primary statistic of interest is that of the Maximum Likelihood Estimate (MLE), or mode in a Bayesian context, the optimal set for the driving values $\boldsymbol{\theta}$ is the MLE (or mode). This is cumbersome since, the best proposal we can chose depends on the answer. We will address this problem through an iterative mode searching algorithm in the next section.

4.3.1 Mode Searching

We propose a method for finding the mode $\hat{\boldsymbol{\theta}} = (\hat{N}_0, \hat{N}_1, \hat{N}_2, \hat{\lambda})$ of the likelihood function. Our method relies on a two level search procedure. The first procedure retrieves seeding values for a second stage steepest descent algorithm. Before the algorithm begins, we initialize the driving set to be $\boldsymbol{\theta}^{(*)}$.

4.3.2 Search Engine

The first level of the search generates points at a random location in the parameter space $\boldsymbol{\theta} = (N_0, N_1, N_2, \lambda) = (\theta_1, \theta_2, \theta_3, \theta_4)$. Likelihood estimates are compute at each of these locations in the space. Hence we have a tuple of the form $(\boldsymbol{\theta}, \tilde{L}_{\boldsymbol{\theta}^{(*)}}(D|\boldsymbol{\theta}))$, where D is our observed data and $\tilde{L}_{\boldsymbol{\theta}^{(i)}}(D|\boldsymbol{\theta})$ is the approximation to the likelihood, where $\boldsymbol{\theta}^{(*)}$ is the current driving set. For some fixed tolerance τ, we accept this randomly generated point $\boldsymbol{\theta}$ if

$$\tilde{L}_{\boldsymbol{\theta}^{(*)}}(D|\boldsymbol{\theta}) > \tau. \tag{4.33}$$

The tolerance should be set high enough so that the threshold is discriminatory, however it must not be set so high that all $\boldsymbol{\theta}$ fail to meet the criterion. A parameter set such that equation (4.33) holds is passed to the second phase of the search procedure.

4.3.3 Steepest Descent

The steepest descent algorithm is launched when a seeding value is found by the searching engine, it is passed to the slave code which performs a local steepest algorithm starting from the seeding value. To find a searching direction, at each iteration (i), for each parameter $\theta_j^{(i)}$ (the j^{th} component of $\boldsymbol{\theta}$) we examine the the point

$$\theta_{K_j}^{(i)} = \theta_j^{(i)} + K\epsilon$$

where $K \in \{-1, 0, +1\}$. We select the search direction Δ from

$$\Delta = \max_{K_1,K_2,K_3,K_4} (\theta_{K_1}^{(i)}, \theta_{K_2}^{(i)}, \theta_{K_3}^{(i)}, \theta_{K_4}^{(i)}).$$

Along the search direction Δ, we move each parameter θ_j according to some fixed step size ϵ, and set

$$\boldsymbol{\theta}_N^{(i)} = \{\theta_1^{(i)} + N\epsilon\delta_1, \theta_2^{(i)} + N\epsilon\delta_2, \theta_3^{(i)} + N\epsilon\delta_3, \theta_4^{(i)} + N\epsilon\delta_4\}$$

where $N = 0, 1, 2, \ldots, Z$ and Z is some predetermined maximum for the number of steps, and $\delta_j \in \Delta$. From this we take the move to be

$$\tilde{\boldsymbol{\theta}} = \max_N \boldsymbol{\theta}_N^{(i)}.$$

The above procedure is done using $\boldsymbol{\theta}^{(i)}$ as the driving values in the proposal distribution for the IS routine.

4.3.4 Validation

Since, at every step, the likelihood is being generated under the importance sampling approxiation (4.13), we may have generated a higher likelihood candidate

point $\tilde{\boldsymbol{\theta}}$ simply by chance. Hence, once a proposed move $\tilde{\boldsymbol{\theta}}$ is found it must be validated as an acceptable move. Our criterion for acceptance of the proposed point, is that it must yield a higher likelihood value by using the values $\tilde{\boldsymbol{\theta}}$ as driving values in the proposal distribution in the IS routine, than was obtained in finding $\tilde{\boldsymbol{\theta}}$ under the driving values $\boldsymbol{\theta}^{(i)}$. If the criteria is met, we set

$$\boldsymbol{\theta}^{(i+1)} = \tilde{\boldsymbol{\theta}}$$

else,

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}.$$

The complete algorithm follows as under the two stage procedure follows below

Algorithm 4.3.1: SEARCH-ENGINE(τ)

for $i \leftarrow 0$ **to** n
  $\theta_1 \sim U(\alpha_1, \beta_1)$
  $\theta_2 \sim U(\alpha_2, \beta_2)$
  $\theta_3 \sim U(\alpha_3, \beta_3)$
  $\theta_4 \sim U(\alpha_4, \beta_4)$
  if $L_{\boldsymbol{\theta}^{(*)}}(y|\boldsymbol{\theta}) \geq \tau$
    SEARCH($\boldsymbol{\theta}$)

Algorithm 4.3.2: SEARCH($\boldsymbol{\theta}^{(0)}$)

for $i \leftarrow 0$ **to** n
  $\Delta = \max_{K_1,K_2,K_3,K_4}(\theta^{(i)}_{K_1}, \theta^{(i)}_{K_2}, \theta^{(i)}_{K_3}, \theta^{(i)}_{K_4})$
  for $N \leftarrow 0$ **to** Z
    $\boldsymbol{\theta}^{(i)}_N = \{\theta^{(i)}_1 + N\epsilon\delta_1, \theta^{(i)}_2 + N\epsilon\delta_2, \theta^{(i)}_3 + N\epsilon\delta_3, \theta^{(i)}_4 + N\epsilon\delta_4\}$
    $\tilde{\boldsymbol{\theta}} = \max_N \boldsymbol{\theta}^{(i)}_N$
  if $L_{\boldsymbol{\theta}^{(i)}}(y|\tilde{\boldsymbol{\theta}}) \leq L_{\tilde{\boldsymbol{\theta}}}(y|\tilde{\boldsymbol{\theta}})$
    $\boldsymbol{\theta}^{(i+1)} \leftarrow \tilde{\boldsymbol{\theta}}$
   else $\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^{(i)}$

Generally the seeding values will be perturbed and this procedure will be repeated to guard against settling on local modes.

Once the driving set $\boldsymbol{\theta}^{(*)}$ is determined, we let the final version of the importance sampler, that is the one we'll use for inference, be written as

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= P(D|\boldsymbol{\theta}) \\ &\approx \frac{1}{N}\sum_{t\in S_t}\frac{P(D,t|\boldsymbol{\theta})}{Q(t|D,\boldsymbol{\theta}^{(*)})}, \end{aligned} \tag{4.34}$$

where

$$Q(T|D,\boldsymbol{\theta}^{(*)}) = Q_2(U|D_2,G,\boldsymbol{\theta}^{(*)})Q_1(G|D_1,\boldsymbol{\theta}^{(*)}).$$

4.4 Surface Splines

Our IS procedure approximates the likelihood at only specified points, however we desire a smooth likelihood function for performing parameter inference. Hence given a set of uniformly placed points, we can generate approximations to the likelihood and interpolate between the points to obtain a *meshed* likelihood surface.

High dimensional mesh generation is a laborious procedure, and generally cannot be done for all the parameters simultaneously at a very fine level. When the full joint likelihood function is desired, a coarse grid approximation is generally computed as an approximation to the full surface. Assuming the surface is reasonably well behaved and the likelihood is computed on a grid at least fine enough to capture the true behavior of the complete surface, interpolating splines can be applied to impute the surface to the level of desired refinement. We accomplish our interpolation via tensor product cubic spline interpolation with knots computed uniformly on the grid. For the boundary conditions, we use the Not-A-Knot method which requires the third derivatives at the second and second to last break points be continuous.

Once our surface has been fitted via the interpolation scheme described about, we will have a continuous approximation to the likelihood surface which is what we need for joint parameter inference. Generally, the point estimate of interest is the MLE, for which can can obtain profile likelihood intervals (see Berger *et al.*, 1999). This is described in the next section.

4.5 Interval Estimation

From the interpolated likelihood surface, we are able to construct approximate marginal intervals for the point estimates of interest. Of the possible point estimates, the MLE is of the most appropriate quantities. The interval we will assume for this quantity stems from the asymptotic distribution

$$-2\log(\lambda) \sim \chi^2_1, \tag{4.35}$$

where $\lambda = L_0(\hat{\theta}_j)/L_1$. $L_0(\hat{\theta}_j)$ is the maximum likelihood estimate under the hypothesized quantity $\hat{\gamma}$, of one of margins, and L_1 is the maximum likelihood estimate under the unconstrained parameter space.

By inverting the expression in equation (4.35), we can obtain an interval for each margin $\hat{\theta}_j$. By conditioning each marginal interval on the maximum likelihood estimates at the remaining parameters, we obtain a set of marginal profile likelihood intervals. These profile likelihood intervals for the marginal of interest are computed as the set of θ_j values such that

$$\log(L_1(\theta_j)) + x_\alpha/2 < log(L_0(\theta_j)$$

where α is the level of the interval.

4.6 Discussion

In this chapter, we have laid out a full importance sampling framework for inferring about the demographic parameter set $\boldsymbol{\theta}$. Key contributions to this method are the determination of model based proposal distributions, which relate to the coalescence process described in chapter 3. In building these proposal distributions, special attention was paid to sampling topologies, such that they provide lineages which are consistent with the mutational type patterns observed in the data (section 4.2.1).

This importance sampling method will generate reliable approximations to the probability of the sequence data under a fixed parameter set $\boldsymbol{\theta}$. Using surface splines (section 4.4) on a grid of specified parameter values ($\boldsymbol{\theta}$) and their corresponding likelihoods, we are able to construct a continuous approximation to the likelihood function. From this, MLEs and their corresponding interval estimates (section 4.5) can be determined.

The next section shows how this method performs on a data set of closely related members of the *Drosophila* family. Accuracy, computation time and limitations will be discussed.

Chapter 5

Case Study: A Recent Divergence In Time Of Closely Related Members Of The Drosophilia Family

The Drosophila family is a closely related family of fruit fly species. Machado *et al.* (2002) has provided a data set for the Drosophila species: *D. persimilis*, *D. p. pseudoobscura*, and *D. p. bogotana*. In this case study, we study the DPS2002 region of the genome, which is common to all of the Drosophila species. The DPS2002 region in Drosophila is of interest since it has undergone a paracentric inversion sometime in its evolutionary past (see section 3.2.7 for details). Laboratory experiments have shown that the DPS2002 region in *D. persimilis* is inverted with respect to the genomes of both *D. p. pseudoobscura* and *D. p. bogotana*.

Inversions provide valuable information about the time since speciation. This is due to the fact that inversions regulate the existence of a fixed-absent lineage in the evolutionary history, since inversions must be fixed in one population and absent from the other. This additional information provides information about the topological structure of the genealogy, from which the sample has descended,

and eliminates uncertainty.

In this case study, we will estimate the species divergence time between *D. persimilis* with *D. p. pseudoobscura* and *D. persimilis* with *D. p. bogotana*, respectively. From 13 sequences from the *D. persimilis* and *D. p. bogotana* groups, and 19 sequences from *D. p. pseudoobscura*, the mutational count summary as described in 3.1 are shown in table 5.1.

Count	Group 1	Group 2	Dpe^1/Dpp^2	Dpe/Dpb^3
n_1	segregating	absent	16	16
n_2	fixed	absent	5	6
n_3	absent	segregating	65	18
n_4	absent	fixed	0	2
n_5	segregating	segregating	0	0
n_6	fixed	segregating	1	0
n_7	segregating	fixed	0	0

[1] *D. persimilis*, 13 sequences
[2] *D. p. pseudoobscura*, 19 sequences
[3] *D. p. bogotana*, 13 sequences

Table 5.1: Segregating sites data collected from members of the Drosophila species.

As described by the topological partition on the space of trees (section 4.2.1), after observing the mutational *type* data, we know which partition of tree space explains the data. The data set for the *Dpe*/*Dpb* groups have the fixed-absent and absent-fixed mutation types, so in our importance sampler, we constrain the proposal of trees to have the fa/af topology as shown in figure 5.1. Also, for the

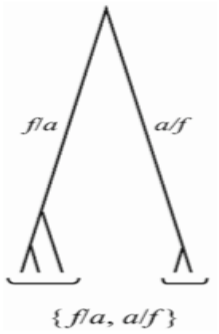

Figure 5.1: Topology of an evolutionary tree with descendants which are fixed and absent in both groups.

data set for the *Dpe*/*Dpp* groups, the mutational types fixed-absent and fixed-segregating are both present. Hence, we constrain the proposed topology to live in the partition with fa/fs lineages, which is shown in figure 5.2.

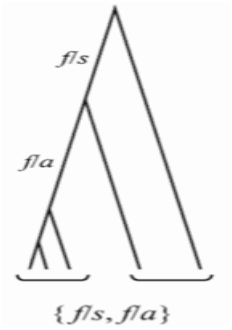

Figure 5.2: Topology of an evolutionary tree with descendants which are fixed and absent in one group and fixed and segregating in the other.

In section 4.2.5, for the basic speciation model, we derived the proposal weights

$$T_{\mathcal{P}(\boldsymbol{S}_L)}(L_1(\boldsymbol{S}_i), L_2(\boldsymbol{S}_i), L_3(\boldsymbol{S}_i))$$

in the product

$$Q_1(D_1, G|\boldsymbol{\theta}^{(*)}) = \prod_{i=L}^{2} T_{\mathcal{P}(\boldsymbol{S}_L)}(L_1(\boldsymbol{S}_i), L_2(\boldsymbol{S}_i), L_3(\boldsymbol{S}_i)) \times P(\boldsymbol{S}_{i-1}|\boldsymbol{S}_i, \boldsymbol{\theta}^{(*)}),$$

where $\boldsymbol{\theta}^{(*)}$ is the driving set (see section 4.3), and $\mathcal{P}(\boldsymbol{S}_L) \in \{\text{ancestral, extant}\} = \{a, e\}$ represents the phase the state currently resides in and $L_i(\boldsymbol{S}_i)$, for $i = \{1, 2, 3\}$, is the number of lineages of type 1,2 and 3 respectively.

These weights are derived such that the generated trees come from the coalescence process, but are conditional on having a topology consistent with the data. Since the process with an inverted region undergoes a slightly modified coalescence process, these weights should be modified as well. We will illustrate this modification on for the fa/af topologies.

5.0.1 Modified Proposal Mechansim

For the model with a paracentric inversion (section 3.2.7) and lineages, we derive the weights

$$T_{\mathcal{P}(\boldsymbol{S}_L)}(L_1(\boldsymbol{S}_i), L_2(\boldsymbol{S}_i), L_3(\boldsymbol{S}_i)),$$

for enforcing topologies with fixed-absent lineages in both groups.

As illustrated in 4.2.5, the boundary condition

$$T_e(i, j, k) = T_a(i, j, k) = 0 \qquad \text{if } k > 0 \text{ and } i, j > 1,$$

states that all lineages of type 1 and 2 must coalesce prior to the formation of a type 3 lineage. The recursive form of $T_e(i,j,k)$ follows as

$$T_e(i,j,0)\left(\frac{\binom{i}{2}}{N_1}+\frac{\binom{j}{2}}{N_2}+\lambda\right)=T_e(i-1,j,0)\frac{\binom{i}{2}}{N_1}+T_e(i,j-1,0)\frac{\binom{j}{2}}{N_2}+T_a(i,j,0)\lambda,$$

which is exactly the same as shown in equation (4.16). It is only the recursion for $T_a(i,j,0)$, which resides in the ancestral phase, that undergoes a modification. This follows as

$$T_a(i,j,0)\left(\frac{\binom{i}{2}}{pN_0}+\frac{\binom{j}{2}}{(1-p)N_0}\right)=T_a(i-1,j,0)\frac{\binom{i}{2}}{pN_0}+T_a(i,j-1,0)\frac{\binom{j}{2}}{(1-p)N_0},$$

where p is the frequency of inverted individuals in species group 1. This modified recursion simply reflects that when there are i lineages left from subgroup 1, the coalescence rate is $\binom{i}{2}/pN_0$. Similarly, when there are j lineages left from subgroup 2, the coalesnce rate is $\binom{j}{2}/(1-p)N_0$.

It should also be noted that a modification in the expected time spent on each level also occurs in the ancestral phase. The expected waiting time spent on each level effects our proposal distribution for applying the mutations on a consistent topology in the ancestral phase. That is, in the 7 dimensional multinomial distribution, shown by equation (4.32), the weights follow as

$$w_l=\frac{1}{\binom{i}{2}/pN_0+\binom{j}{2}/(1-p)N_0}.$$

The proposal distribution for the three remaining topologies undergo similar modifications. These modifications are fully documented in Leman *et al.* (2005).

5.1 Mode searching

We performed the mode searching algorithm (section 4.3.1) on both data sets shown in table 5.1, in order to obtain the driving set $\boldsymbol{\theta}^{(*)}$ for the proposal mechanism. It was found that the MLE of the inversion frequency p was at a very small value (≈ 0.00001) for both data sets. This result is consistent with the findings of Dobzhansky and Powell (1975). The likelihood curve for p (conditioned at the MLEs $\{u\hat{N}_0, u\hat{N}_1, u\hat{N}_2, \lambda/u\}$) shows very flat profiles, suggesting that it does not influence the estimates of the other parameters. For this reason, we arbitrarily fixed the inversion frequency ($p = 0.00001$) for the remainder of the analysis. The entire mode searching process for each data set required approximately 2 weeks of time to perform.

5.2 *D. p. bogotana and D. persimilis* Study

For the *Dpe/Dpb* data set (column 5 of table 5.1), we wish to test the hypothesis that the three population sizes are equal. The null (H_0) and alternative (H_a) hypotheses follow as

$$\begin{aligned} H_0 &: N_0 = N_1 = N_2 \\ H_a &: N_i \neq N_j \qquad \text{for some } i \text{ and } j \in \{1,2,3\}. \end{aligned} \tag{5.1}$$

Table 5.2 shows the maximum likelihood estimates for the parameters $\{u\hat{N}_0, u\hat{N}_1, u\hat{N}_2, \lambda/u\}$ under the unconstrained and constrained hypotheses.

	parameter	Unconstrained	Constrained (H_0)
	λ/u	0.17	0.18
	uN_0	2.31	3.21
	uN_1	2.91	3.21
	uN_2	3.51	3.21
Likelihood		1.05×10^{-5}	8.78×10^{-6}
P-Value			0.83

Table 5.2: *D. p. bogotana* and *D. persimilis* parameter inferences

From the likelihood ratio test, we are able to test hypothesis 5.1. The last column of table 5.2 shows the P-value under the likelihood ratio test and determines that null hypothesis in (5.1) cannot be rejected.

5.2.1 Analysis Of The Importance Sampler

For the *Dpe/Dpb* data set, we set out to determine the accuracy in the importance sampling technique. We've calculated exact conditional likelihood curves for the λ/u parameter. The remaining parameters $\{uN_0, uN_1, uN_2\}$ have been fixed to

their unconstrained MLEs (see table 5.2). From the exact recursion method, under the gene incompatibility model 3.2.7, we have generated the exact likelihood curve. This is illustrated by the red curve in figure 5.3. The importance sampler was executed using 500,000 proposed sample genealogies and an approximate (conditional) likelihood curve was drawn. This was repeated 18 times, for a total of 9×10^6 sample genealogies. The individual IS calculated likelihood curves (500,00 genealogies) are plotted in grey, whereas the average of all of these (9×10^6 genealogies) is plotted in blue in figure 5.3.

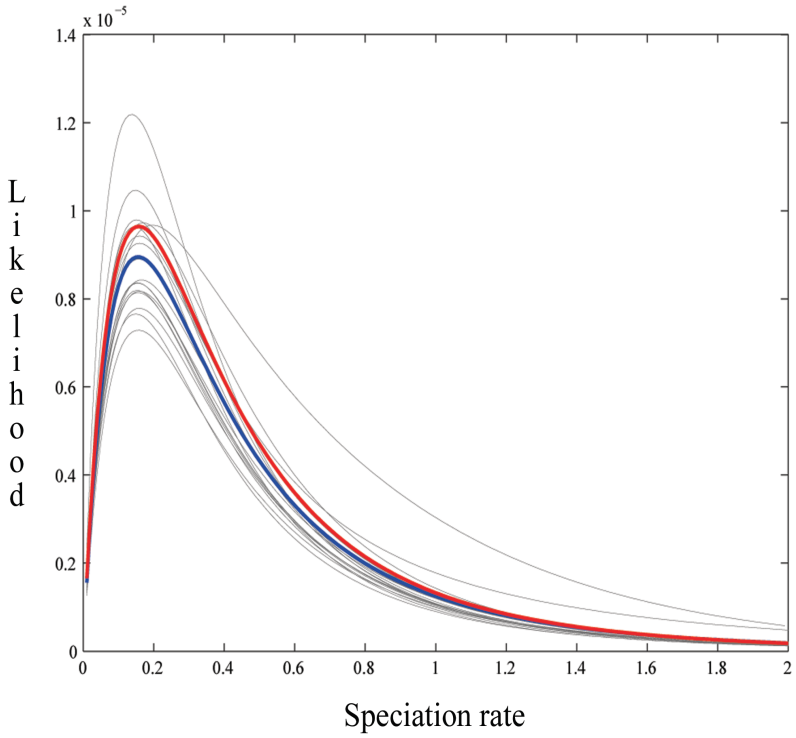

Figure 5.3: Conditional likelihood curves for the λ/u parameter. Grey curves represent IS approximations each using $500,000$ genealogies. The blue curve is the average of all grey curves (9×10^6 genealogies). The red curve represents the true likelihood curve.

From figure 5.3, the modal height of the averaged IS curve (blue) appears to be

somewhat lower than the modal height under the exact recursion. However, this is not necessarily problematic since the heights of likelihood curves are irrelevant for inferential purposes. What is relevant is if this height difference is uniform over the whole range of likelihood values.

Figure 5.4 examines the differences between the approximated log-likelihood curve using 9×10^6 genealogies and the exact log-likelihood curve, where we have scaled both curves to their maximal values.

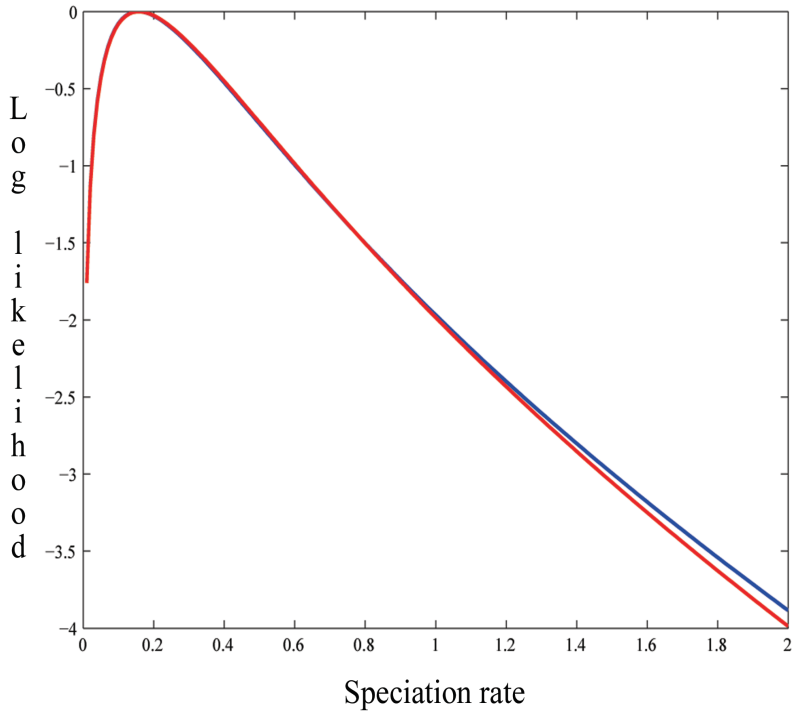

Figure 5.4: Scaled log likelihood comparison between IS approximated curve and the exactly computed curve.

Figure 5.4 shows that the IS approximated likelihood curve compares well to the exact curve, in both shape and location of the MLE. We also notice that the ap-

proximated curve becomes wider than the exact curve as the range deviates from the driving set. However, this deviation is hardly detectable in the range of 2 log-likelihood units which corresponds to the 95% confidence region.

5.2.2 Computational Demands

Each of the curves in figures 5.3 and 5.4 was computed using 200 points in the range of $(0, 2]$. Each of these points computed under the exact recursion required $\sim$4 hours on a Macintosh PowerPC G5 (2.5-GHz processor, 3.5 GB DDR SDRAM). The full exact likelihood curve (red) required $\sim$800 computing hours. For each of the importance sampling curves, based on 500,000 genealogies, $\sim$30 minutes was required in computing time.

5.3 *D. p. pseudoobscura and D. persimilis* Study

The data set *Dpe/Dpp* (column 4 of table 5.1) is much more computationally demanding than the previous data set. Due to the additional samples (13 and 19 sequences respectively), the number of tree topologies present in tree space is much larger than in the *Dpe/Dpb* case (13 sequences each). Beyond this, we also notice that there is an increase in the number of mutations. Under the exact recursion, the computational complexity increases multiplicatively as the number of mutations increases in each group. This creates a massive computational strain when using the exact recursion. Under the IS method, increasing the number of mutations requires virtually no more computational effort than smaller numbers of mutations. We independently examine the three hypotheses $N_0 = N_1$, $N_0 = N_2$ and $N_1 = N_2$. The results found under the IS algorithm are shown in table 5.3

The results in table 5.3 were calculated using 4×10^6 genealogies. From these

	parameter	Unconstrained	$N_0 = N_1$	$N_0 = N_2$	$N_1 = N_2$
	λ/u	0.12	0.12	0.12	0.09
	uN_0	0.81	2.51	11.51	0.91
	uN_1	2.71	2.51	2.71	12.61
	uN_2	18.21	15.61	11.51	12.61
Likelihood		4.59×10^{-6}	2.88×10^{-6}	2.34×10^{-7}	9.23×10^{-9}
P-Value			0.33	1.5×10^{-2}	4.3×10^{-4}

Table 5.3: *D. persimilis* and *D. p. pseudoobscura* parameter inferences

results, the only hypothesis that we fail to reject is $N_0 = N_1$, suggesting that the population of *D. p. pseudoobscura* is significantly larger than the other populations in the model.

Under the maximum likelihood framework, confidence intervals for the parameter estimates found in table 5.3 are found using the profile likelihood (see section 4.5). Estimation of these intervals generally requires a full likelihood surface, which we generate through a mesh refinement using interpolating splines which is described in section 4.4. Profile likelihood curves, computed from the full 4-dimensional likelihood surface, for the parameters uN_1 and λ/u are shown in figure 5.5.

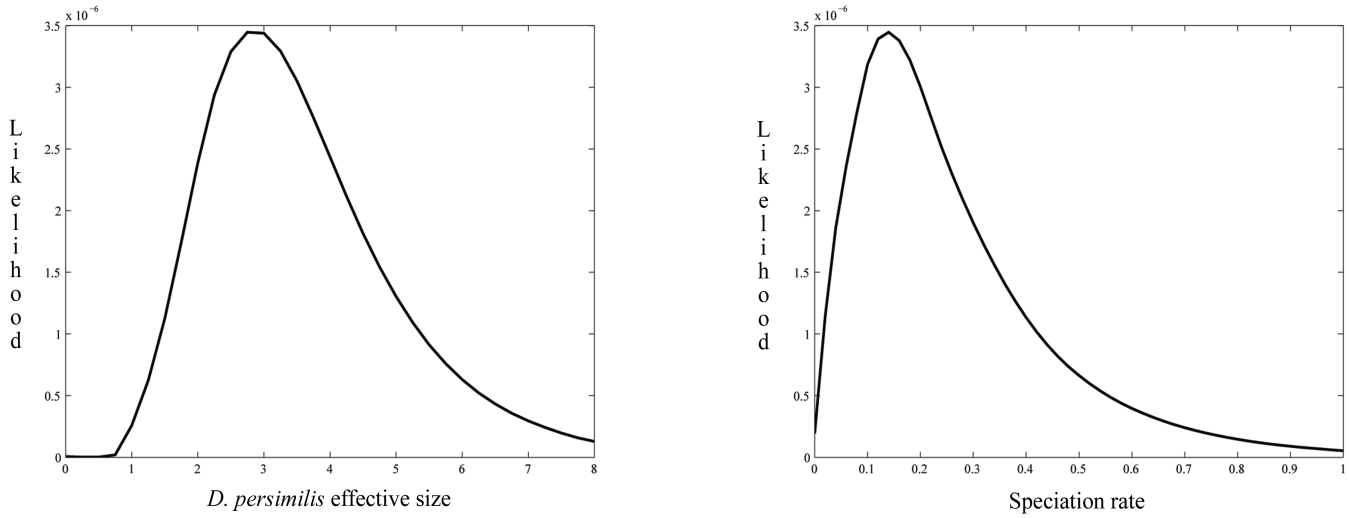

Figure 5.5: Profile likelihood curves for the effective population size of *D. persimilis* and the divergence time between the groups *Dpe*/*Dpp*

Table 5.4 summarizes the 90% marginal confidence intervals for each of the population parameters. Each of the estimates in table 5.4 was based on 4×10^6 genealogies. The likelihood corresponding to the effective population size (uN_2) of the *D. p. psuedoobscura* group has a very heavy right tail, which limits the computation of the right confidence bound for the parameter. In chapter 6, we will study an MCMC procedure for computing the full 4-dimensional posterior distribution, under which, such boundary limitations will be avoided.

parameter	MLE	90% confidence intervals
λ/u	0.12	$(0.02, 0.46)$
uN_0	0.81	$(0.1, 5.0)$
uN_1	2.71	$(1.25, 5.75)$
uN_2	18.21	$(8, -)$

Table 5.4: Interval estimates for *D. persimilis* and *D. p. psuedoobscura*

5.4 Calibrating By Mutation Rate

Up until now, we have scaled all of the parameters to the mutation rate u. While this parameter is not estimable under a coalescence model, it can be obtained under experimental means and is necessary for direct inference on the un-scaled parameters $\{N_0, N_1, N_2, \lambda\}$.

5.4.1 Population Sizes

A coalescence event occurs at rate $\frac{1}{N_i}$, for $i = (0, 1, 2)$, depending on subgroup, hence the ratio

$$\frac{u}{1/N_i} = uN_i,$$

a ratio of rates, is measured on the unit scale $\frac{\text{mutations}}{\text{coalescence}}$. The mutation rate u is defined on the scale of mutations/year. However, in practice, we will obtain rates $\hat{u}$ that are scaled as mutations/kb/year, where kb denotes *kilo bases*. We related the effective population size with half the inverse rate of coalescence (see Slatkin, 1991). Under the assumption of 4 generations per year (Schaeffer, 1995) (which is specific for Drosophila species), we scale the estimated quantity uN_i as

$$4 \times uN_i \times \frac{1}{2\hat{u}} \frac{1}{\mathcal{K}}, \tag{5.2}$$

where $\mathcal{K}$ denotes the number of kb. After the scaling in (5.2), the units become $\frac{\text{generations}}{\text{coalescence}}$, which is the measurement unit for effective population size. This measures the rate of reproduction between coalescence events and is thus identified with the population size.

5.4.2 Divergence Times

In order to better understand the estimate of the scaled speciation rate parameter λ/u, we transform to a yearly time scale. Recall that λ is the rate of speciation under an exponential waiting time model. Hence, the parameter $1/\lambda$ relates to the expected time until speciation. After scaling our estimate to kb, we obtain the estimate

$$\frac{u}{\lambda} \times \frac{1}{\hat{u}\mathcal{K}}, \tag{5.3}$$

which will produce an estimate of time in years.

5.4.3 Discussion

Hey and Nielsen (2004) estimate the mutation rate as 5.3×10^{-6} mutations/kb/year from an analysis across 14 regions for the Drosophila species *D. persimilis* and *D. p. psuedoobscura*. Tamura *et al.* (2004) estimate this same mutation rate as 1.1×10^{-5} after adjusting for codon usage bias. We investigate the population sizes and divergence rates under both estimates. These are listed in table 5.5. These

parameter	HN[a]	TSK[b]
Divergence Time	1.76	0.85
Ancestral effective size	0.34	0.17
D. persimilis effective size	1.15	0.55
D. p. psuedoobscura effective size	7.70	3.71

All estimates $\times10^6$
[a] $\hat{u} = 5.3 \times 10^{-6}$
[b] $\hat{u} = 1.1 \times 10^{-5}$

Table 5.5: Adjusted divergence and population size estimates for *D. persimilis* and *D. p. psuedoobscura*

results suggest a divergence time between *D. persimilis* and *D. p. psuedoobscura* of 850 thousand years. This result is identical to that obtain by Tamura *et al.* (2004) based on a genome wide experiment.

Based soley on the DPS2002 locus, Hey and Nielsen (2004) estimate the parameters $(\lambda/u = 0.15, N_0 = 1.1, N_1 = 2.4, N_2 = 21.6)$ which correspond well to the estimates found under our method $(\lambda/u = 0.12, N_0 = 0.81, N_1 = 2.71, N_2 = 18.21)$. In fact under a likelihood ratio test, the discrepancies in these estimates are insignificant, adding credibility to the methodology.

Chapter 6

Good Chains From Bad Proposals: The Evolutionary Forest Algorithm

Up until this point, we have described a general class of models which can be used to infer about aspects of the evolutionary process. Parameters which pertain to these processes include the time since divergence, effective population sizes, and migration rates (see chapter 3). The general modeling setup which we have developed in chapter 3 gives rise to the joint probability of the data and the genealogy t. However, for inferential purposes, we are interested in the marginal distribution

$$P(D|\boldsymbol{\theta}) = \sum_{t \in \Omega_T} P(D, t|\boldsymbol{\theta}) \tag{6.1}$$

where $\boldsymbol{\theta}$ incorporates the set of population parameters of interest. In chapter 4, we proposed an importance sampling scheme which aided us in computing the probability at a single point of the parameter space $\boldsymbol{\theta}$. Unfortunately, in the context of the case study (see chapter 5), the full likelihood surface was necessary for parameter inference, however exact computation of this is infeasible. Beyond the computational burden of likelihood construction, the proposal mechanisms

used in the importance sampler are dependent on a single fixed set of driving parameter values ($\boldsymbol{\theta}^{(*)}$) which require considerable tuning using the mode searching algorithm presented in section 4.3.1.

In this chapter, we will develop a Markov Chain Monte Carlo (MCMC) method for estimating the population parameters under the divergence model given in section 3.2.1. While we specify the simple model for general purposes, modifications to the gene incompatibility model (section 3.2.7) and other related models are straightforward.

Our MCMC algorithm will ultimately yield the joint posterior distribution ($p(\boldsymbol{\theta}|D)$) of interest, where $\boldsymbol{\theta} = \{\lambda/u, uN_0, uN_1, uN_2\} = \{\Lambda, \theta_0, \theta_1, \theta_2\}$. Recall, the relationship between the divergence time τ and the speciation rate λ is $E[\tau] = \lambda^{-1}$.
Inherent to the problem of estimating the population parameters is the problem of constructing the phylogenetic tree which represents the evolutionary history of the gene sample. While MCMC algorithms provide a powerful frame work for reconstructing complicated posterior distributions, proposal distributions used in moving through the targeted space are a key concern and can require special attention.

Simple proposal mechanisms used for updating genealogies based on local moves generally convergence poorly, since tree space is not well connected. To aid in moving around the space of genealogies, we will develop an augmentation of the tree space which will improve the mixing properties of the Markov chain. Instead of using a single genealogy, we will redefine the probability distribution which is

simultaneously comprised of multiple genealogies. While we change the probability distribution which pertains to the space of trees, the parameter space ($\boldsymbol{\theta}$) will be unaffected.

This method greatly reduces the human efforts required by our importance sampler, and yields massive time reductions over methods in which the exact recursion probabilities are used (see section A.11).

6.1 Proposal Distributions

Before discussing the EF algorithm, we analyze the proposal distributions outlined in chapter 4. While these proposal distributions were derived using insights from the coalescence model, this analysis highlights the inherent difficulties in proposing sample draws from the coalescent.

Since ideal proposal distributions should resemble the target distribution, it is important to measure how close these distributions are to each other. One quantitative method for assessing the proposals performance is by analyzing the importance weights

$$w(x) = p(x)/g(x),$$

where we have used $g(x)$ to symbolize the proposed density (or mass) value at x, and $p(x)$ is the density (or mass) under the true model. Clearly, if $g(x) = p(x)$, then $w(x) = 1$, for all values of x. Letting $x_i \sim g(x)$ for $i = (1, \ldots n)$, we can construct n importance weights, which we denote w_i. A summary of how variable these weights are is described by the *coefficient of variation*

$$c_\nu = \frac{\sum_{i=1}^{n} (w_i - \bar{w})^2}{(n-1)\bar{w}^2} = \left(\hat{\sigma}/\hat{\mu}\right)^2,$$

where $\bar{w} = \sum_{i=1}^{n} w_i/n$. This coefficient of variation gives a measure of exactly how variable the values of w_i are, relative to its mean. Generally speaking, the lower the coefficient of variation, the better the proposal distribution matches the true target distribution. For example, if $g(x) = p(x)$, then $c_\nu = 0$.

Another quantity of interest, when analyzing proposal distributions, is the effective sample size (ESS). This is given by

$$\text{ESS(n)} = \frac{n}{1 + c_\nu},$$

where n is the quantity of sampled values. Roughly speaking, this quantity measures how many exact samples (generated under $p(\cdot)$) our proposed samples (generated under $g(\cdot)$) are worth. Notice when $c_\nu = 0$, ESS(n)=n, so the effective sample size is equal to the true number of samples. If c_ν is large, then the effective sample size will be reduced. In chapter 4, section 4.2, we considered a set of proposal distributions that were used for construction of the genealogy. This was a two tier process. The first tier constructed the tree topology via the coalescence process, and the second tier *sprinkled* mutations over the tree structure using weighted multinomial distributions on each mutation type. While this is a seemingly reasonable method for constructing the proposed genealogy, it can at times yield probability weights that greatly differ from the true quantities.

We examine the quantities c_ν and ESS(n) under the proposal mechanism studied in 4.2, both with and without the recursion constraints discussed in section 4.2.1. For the data sets given in 5.1, we sampled $1,000,000$ genealogies under the proposal distribution, with driving values set to the MLE found in chapter 5 (tables

5.3 and 5.2), and computed the importance weights for each sampling. Table 6.1 summarizes the results of both ESS and c_ν for the data in table 5.1.

Data set	ESS$(10e+6)^\dagger$	$c_\nu^\dagger$	ESS$(10e+6)^\ddagger$	$c_\nu^\ddagger$
Dpe/Dpp	9.6	7.2489e+03	42.3	2.3658e+04
Dpe/Dpb	137.9	1.0421e+05	561.7	1.7793e+03

† Without topology constraints

‡ With topology constraints

Table 6.1: Expected sample sizes and coefficients of variation from proposal distributions, based on $1,000,000$ samples.

From the comparison of the data sets (*Dpe/Dpp* with 32 individuals and *Dpe/Dpb* with 26 individuals), we notice that the larger data set has a far lower sampling efficiency, as determined by the ESS. This is due to the fact that the larger data has a vastly larger tree space. Generally larger data sets have poor sampling efficiency due to the stochastic variability being so much higher. This presents a dilemma since, more data suggests more accuracy in parameter estimation, but more data enlarges the tree space and creates a high degree of sampling variation which in turn hinders the estimation.

We note that by using the topology constraints (section 4.2.1), we obtain approximately a 4 fold increase in sampling efficiency in both data sets. However, even though the sampling performance is increased by adding the constraints, which ensure consistent trees, the performance is still quite poor by most standards.

Instead of developing a new set of proposal distributions, we will explore a new MCMC methodology that helps to *uniformize* the probability distribution on the tree space and ultimately improve sampling performance.

6.1.1 Gibbs sampling

Before introducing the EF algorithm, we first describe a more theoretically straightforward method to illuminate the basic idea in EF. One way to avoid the summation in (6.1) is to consider the joint posterior distribution $p(\boldsymbol{\theta}, t|D)$, whose marginal distribution $p(\boldsymbol{\theta}|D)$ is the target distribution. Sampling from the posterior distribution $p(\boldsymbol{\theta}, t|D)$ can be accomplished by the use of a Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) with Metropolis-Hastings steps. That is, we sample iteratively from the distributions

$$p(\boldsymbol{\theta}|t, D) \quad \text{and} \quad p(t|\boldsymbol{\theta}, D).$$

Iterating through this algorithm a sufficient number of times yields samples from the joint distribution $p(\boldsymbol{\theta}, t|D)$. Albeit, this method is conceptually straightforward, convergence in the tree space is usually problematic. The main difficulty in this sampling approach is proposing updates to the tree history which explore the tree space efficiently and have good mixing properties.

For this reason, we introduce a new sampling algorithm called the Evolutionary Forest (EF).

6.2 The Forest Approach

In this thesis, we contribute a new MCMC algorithm which requires simple proposals, yet achieves adequate mixing and convergence by augmenting the space of trees with a collection of genealogies (forest). Instead of considering the joint distribution $p(\boldsymbol{\theta}, t|D)$, we will augment the dimension in which $t \in \Omega_T$ lives, such that the marginal distribution of $\boldsymbol{\theta}$ is invariant under the augmentation. Another feature of the augmentation is that the Markovian type moves in the MCMC algorithm will be simplified. We start by defining the forest sample space by

$$\Omega_F = \left\{ \biguplus_{i=1}^{K} t_i : \ t_i \in \Omega_T \right\}, \tag{6.2}$$

for K a fixed number of trees. The operation $\biguplus$ denotes a multiset union of trees, where a multiset is an unordered set for which multiplicity is preserved. For example, let $A = \{t_1, t_2, t_3\}$ and $B = \{t_1, t_4\}$. Under the operation $\uplus$, we have $A \uplus B = \{t_1, t_1, t_2, t_3, t_4\}$. For each forest $f \in \Omega_F$, we denote the multiplicity of a tree $t \in f$ as $M(t, f)$. Hence, we allow for multiple copies of the same tree in the forest.

We explicitly assign the probability of the joint posterior distribution of f, $\boldsymbol{\theta}$, and the data D by the system

$$\begin{aligned} q(f, \boldsymbol{\theta}|D) &= C \sum_{i=1}^{K} p(t_i, \boldsymbol{\theta}|D) \\ q(D) &= p(D), \end{aligned} \tag{6.3}$$

where C is a normalizing constant, such that

$\sum_{f \in \Omega_F} \int_{\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}} q(f, \boldsymbol{\theta}|D) d\boldsymbol{\theta} = 1$. Before proceeding, we will derive some useful facts concerning the distribution $q(f, \boldsymbol{\theta}|D)$.

The inverse of the normalizing constant is calculated as

$$\begin{aligned} \frac{1}{C} &= \binom{N+K-1}{K} \frac{K}{N} \\ &= R(N,K) = R \end{aligned} \tag{6.4}$$

where $N = |\Omega_T|$ is the size of tree space. That is, N describes the total number of trees in the space of trees, where we assume $N < \infty$. This normalizing constant is derived from the fact that there are $\binom{N+K-1}{K}$ possible multisets. This is equivalent to the problem of randomly drawing from N indistinguishable balls and placing them into K distinguishable urns. The solution to the number of ways this can be accomplished is $\binom{N+K-1}{K}$ ways (Ross, 2005). Each forest has K positions for trees to reside in, hence we can consider the forest space as being a partition on $\binom{N+K-1}{K} \times K$ positions. By symmetry, each tree appears in the fraction $\frac{1}{N}$ of the $\binom{N+K-1}{K} \times K$ positions, which results in (6.4). Another useful characterization of R is

$$\sum_{f \in \Omega_F} \sum_{t \in f} (\cdot) = R \sum_{t \in \Omega_T} (\cdot). \tag{6.5}$$

That is, R is the number of times each tree is represented in the full forest space, and by symmetry is equal for every $t \in \Omega_T$.

We also have that $q(f|\boldsymbol{\theta}, D) = C \sum_{t \in f} p(t|\boldsymbol{\theta}, D)$ forms a proper distribution since

$$\begin{aligned} 1 &= \sum_{f \in \Omega_F} \int q(f, \boldsymbol{\theta}|D) d\boldsymbol{\theta} = \sum_{f \in \Omega_F} \int C \sum_{t \in f} p(t, \boldsymbol{\theta}|D) d\boldsymbol{\theta} \\ &= \int \left(\sum_{f \in \Omega_F} \sum_{t \in f} C p(t|\boldsymbol{\theta}, D) \right) p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}. \end{aligned}$$

From equation 6.5, we can write the above term in parentheses as

$$\begin{aligned} \sum_{f \in \Omega_F} \sum_{t \in f} C p(t|\boldsymbol{\theta}, D) &= R \sum_{t \in \Omega_T} C p(t|\boldsymbol{\theta}, D) \\ &= RC = 1. \end{aligned}$$

Hence, $\sum_{f \in \Omega_F} q(f|\boldsymbol{\theta}, D) = C \sum_{f \in \Omega_F} \sum_{t \in f} p(t|\boldsymbol{\theta}, D) = RC = 1$ and holds for all $\boldsymbol{\theta}$. From this, we see that the definition of the distribution of $f \in \Omega_F$ given in equation (6.3) adheres to familiar rules of probability distributions. That, is jointly and conditionally, the total probability mass of Ω_F is 1.

The following theorem is the key to why we can use the augmentation defined in equation (6.3) and still perform inference on the parameter set $\boldsymbol{\theta}$.

Theorem 1. *The marginal posterior distribution of* $\boldsymbol{\theta}$ *is invariant under the forest augmentation defined by equation* (6.3).

Proof. The marginal posterior distribution is

$$\begin{aligned} q(\boldsymbol{\theta}|D) &= \sum_{f\in\Omega_F} q(f,\boldsymbol{\theta}|D) \\ &= \sum_{f\in\Omega_F}\left(\sum_{t\in f} Cp(t,\boldsymbol{\theta}|D)\right) \\ &= p(\boldsymbol{\theta}|D)\sum_{f\in\Omega_F}\left(\sum_{t\in f} Cp(t|\boldsymbol{\theta},D)\right) \\ &= p(\boldsymbol{\theta}|D)\left(\sum_{t\in\Omega_T} RCp(t|\boldsymbol{\theta},D)\right) = p(\boldsymbol{\theta}|D). \end{aligned}$$

Hence, the marginal posterior distribution of $\boldsymbol{\theta}$ under the forest augmentation is the same as under the original tree construction. That is, under the augmentation defining f, the margin on $\boldsymbol{\theta}$ is preserved. □

Notice that implied by $q(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta}|D)$, we have $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, since $q(D) = p(D)$. From the construction for f, we propose to sample from the posterior distribution

$$\begin{aligned} q(\boldsymbol{\theta}, f|D) &\propto q(D, f|\boldsymbol{\theta})q(\boldsymbol{\theta}) \qquad (6.6) \\ &= C\sum_{i=1}^{K} p(D, t_i|\boldsymbol{\theta})p(\boldsymbol{\theta}). \end{aligned}$$

The above summation follows from the definition of the joint posterior probability

on the forest space, since

$$\begin{aligned} q(D, f|\boldsymbol{\theta}) &= q(\boldsymbol{\theta}, f|D)\frac{q(D)}{q(\boldsymbol{\theta})} \\ &= \sum_{t \in f} Cp(\boldsymbol{\theta}, t|D)\frac{p(D)}{p(\boldsymbol{\theta})} \\ &= \sum_{t \in f} Cp(D, t|\boldsymbol{\theta}). \end{aligned}$$

We will see that augmenting the parameter space with a forest of trees increases the sampling efficiency in the MCMC routine, which accelerates convergence to the true distribution. Also, construction of the individual tree proposal distributions can be done in a *naive* manner. That is, we can use simple proposal distributions and still maintain rapid convergence to the target distribution. In particular, we may use an independent proposal distribution on the individual trees rather than proposing new trees based on complicated rules which define local moves.

6.3 The EF Algorithm

Our algorithm relies on Metropolis-Hastings within Random Gibbs Moves (Robert and Casella, 1999). That is, we sample $\nu \sim \text{Bernoulli}(p)$ to choose either a parameter space move or forest move respectively. A reasonable choice for p is $\frac{1}{2}$. Parameter moves are proposed according to a normal random walk, while forest moves are proposed by independent tree proposals.

Explicitly, the algorithm is:

1. set parameter starting values $\boldsymbol{\theta}^{(0)} = \{\Lambda^{(0)}, \theta_0^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}\}$.

2. generate K genealogies $f^{(0)} = \{t_1^{(0)} \dots t_K^{(0)}\}$.

3. For $i = 1, \dots, N$

   Sample $\nu \sim$ Bernoulli(p).

   - if $\nu = 1$ (parameter space move)

     Propose parameter updates $\boldsymbol{\theta}^{(c)} \sim N(\boldsymbol{\theta}^{(i-1)}, \sigma I)$.

     With probability α_p set $\boldsymbol{\theta}^i = \boldsymbol{\theta}^c$, where

$$\alpha_p = \min\left(1, \frac{q(D, f^{(i-1)}|\boldsymbol{\theta}^{(c)})p(\boldsymbol{\theta}^c)}{q(D, f^{(i-1)}|\boldsymbol{\theta}^{(i-1)})p(\boldsymbol{\theta}^{(i-1)})}\right). \tag{6.7}$$

   - if $\nu = 0$ (forest space move)

     Sample $t_j \sim Unif(f^{(i-1)})$, so that $f^{(i-1)} = t_j \uplus f_{-j}^{(i-1)}$.

     Propose a new tree independently from the distribution $g(t^c|\boldsymbol{\theta}^{(i)}, D)$ and denote $f^{(c)} = t^c \uplus f_{-j}^{(i-1)}$. Acceptance of the j^{th} tree occurs with acceptance probability α_f, where

$$\alpha_f = \min\left(1, \frac{q(f^{(c)}|D, \boldsymbol{\theta}^{(i)})M(t^c, f^{(c)})g(t_j|\boldsymbol{\theta}^{(i)}, D)}{q(f^{(i-1)}|D, \boldsymbol{\theta}^{(i)})M(t_j, f^{(i-1)})g(t^c|\boldsymbol{\theta}^{(i)}, D)}\right). \tag{6.8}$$

Continuing this algorithm for a sufficient number of iterations to guarantee convergence results in sample draws from $q(f, \boldsymbol{\theta}|D)$.

6.3.1 Algorithmic Details

In the above algorithm, we iterate between draws from the distributions: $p(\boldsymbol{\theta}|D, f)$ (parameter space) and $p(f|D, \boldsymbol{\theta})$ (a forest of trees). Upon convergence, this yields

sample draws from $q(f, \boldsymbol{\theta}|D)$. In the parameter updates, we use in the acceptance ratio (6.7), the quantity

$$\sum_{k=1}^{K} p(D, t_k^{(i-1)}|\boldsymbol{\theta}^{(c)})p(\boldsymbol{\theta}^c) = q(D, f^{(i-1)}|\boldsymbol{\theta}^{(c)})p(\boldsymbol{\theta}^c).$$

The desired acceptance probability at this step is

$$\frac{q(\boldsymbol{\theta}^c|D, f)g(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^c)}{q(\boldsymbol{\theta}^{(i-1)}|D, f)g(\boldsymbol{\theta}^c|\boldsymbol{\theta}^{(i-1)})} = \frac{q(\boldsymbol{\theta}^c|D, f)}{q(\boldsymbol{\theta}^{(i-1)}|D, f)},$$

where the distribution $g(\cdot|\mu)$ represents a normal distribution centered at μ and cancels out of the ratio since it is a symmetric distribution. The necessary conditional distribution $q(\boldsymbol{\theta}|D, f)$ for this acceptance probability follows as

$$\begin{aligned} q(\boldsymbol{\theta}|D, f) &= \frac{q(D, f|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(f|D)p(D)} \\ &= \frac{\sum_{k=1}^{K} p(D, t_k|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(f|D)p(D)}, \end{aligned}$$

which is known up to the distributions $q(f|D)$ and $p(D)$, where

$$q(f|D) \propto \int_{\boldsymbol{\theta}} q(D, f|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

This integral is difficult to compute, since we must integrate all the parameter dimensions from any arbitrary tree. Fortunately, at each iteration $q(f|D)$ is a constant within each block parameter draw, so these terms cancel in (6.7).

For updating the forest, we update the j^{th} tree, where $t_j \sim Unif(f^{(i-1)})$. Since a tree can exist in the forest with multiplicity, the probability of proposing $f^{(c)}$

from $f^{(i-1)}$ is

$$\frac{M(t_j, f^{(i-1)})}{K} g(t^{(c)}|\boldsymbol{\theta}^{(i)}, D), \tag{6.9}$$

where $g(t|\boldsymbol{\theta}, D)$ is the proposal distribution on trees (see section 4.2). Note that the full conditional forest distribution

$$\begin{aligned} q(f|D, \boldsymbol{\theta}) &= \frac{q(f, D|\boldsymbol{\theta})}{q(D|\boldsymbol{\theta})} \\ &= \frac{\sum_{i=1}^{K} p(t_i, D|\boldsymbol{\theta})}{q(D|\boldsymbol{\theta})} \end{aligned} \tag{6.10}$$

is known up to the distribution $p(D|\boldsymbol{\theta})$, which is constant for fixed $\boldsymbol{\theta}$. Combining equation (6.10) and equation (6.9) yields the importance ratio used in the Metropolis-Hastings acceptance probability, given by equation (6.8).

6.4 Relationship To Parallel Tempering

Methods where multiple chains are run simultaneously (Geyer, 1991) can often increase sampling performance. In parallel tempering, a transformation of the probability distribution through an energy function is performed, such that the distribution is *flattened*, so that movement between regions of high posterior density (or mass) is increased and mixing is improved. Typically, a *temperature ladder* is specified in parallel tempering and sampling efficiency in the chain is maximal when the temperature is *hot*. However, samples drawn from this high temperature chain do not reflect samples that would be drawn from the target chain. Therefore, in order to to retain samples from the distribution of interest, the chain must be cooled. The EF algorithm mimics the basic principle in parallel tempering of *flattening* the space of trees. By increasing the number of trees in the forest space,

the relative probability contribution of each individual tree is diminished, resulting in a distribution on forest space that looks like a flattened tree space. A key advantage of EF over tempering is that the margins in the space of population parameters is retained (see Theorem 1), so there is no need for the additional complexity of multiple chains. Computational complexity is increased only very slightly since K evaluations of tree probabilities must be evaluated in order to determine the probability of each forest. Hence, in the EF algorithm, sampling efficiency is increased with relatively little computational overhead.

6.5 Results and Discussion

A simulation study was performed to assess the convergence of the EF algorithm. In this study, population parameters were uniformly generated with $\Lambda \in (0, 5]$ and $\theta_i \in (0, 30]$ for $i = \{0, 1, 2\}$. Under the simulated parameters, the coalescence process was simulated, starting with 5 samples in each deme, and counts for the 7 joint summary statistics were collected. This process was repeated 100 times, so that each run produced a collection of summary statistics for which we could estimate the population driving values using the EF algorithm. We excluded generated data sets in which fewer than 15 segregating mutations were counted, since data generally has more than 15 segregating mutations. For estimation, prior distributions for the population parameters were chosen to be: $\Lambda \sim \text{Unif}(0, 5]$ and $\theta_i \sim \text{Unif}(0, 30]$ for $i = \{1, 2, 3\}$, which are the same as the simulated values. Estimation of the joint posterior distribution was performed using the EF algorithm with $K = 50$. In all data sets, convergence of the posterior distribution required fewer than 100,000 MCMC iterations, as determined by trace plots. We computed 95% credible intervals for each of the parameters and calculated the frequency of

times each of the intervals covered it's known parameter value. Coverage probabilities for each of the marginal distributions are shown in Table 6.2.

parameter	coverage probability
Λ	0.93
θ_0	0.96
θ_1	0.96
θ_2	0.96

Table 6.2: Coverage probabilities for the 0.95 level credible intervals under simulation.

We found the 95% credible intervals covered the known parameter values with the appropriate frequency, which is indicative that the EF algorithm converges to the correct target distribution.

6.5.1 Comparison To Exact Posterior Calcuations

Using the data set provided by Machado *et al.* (2002), for the *Dpe*/*Dpb* groups, we compare exact marginal distributions, as determined by our exact recursion (see appendix A), to those obtained using EF. We first compare the likelihood curves for the divergence rate Λ conditional on $\theta_0 = 2.31, \theta_1 = 2.91, \theta_2 = 3.51$, the pre-calculated maximum likelihood values found in chapter 5 (table 5.2). For the EF simulations, we have chosen to examine forests of sizes $K = \{1, 10, 25\}$. For this comparison, we specify the diffuse prior distribution $p(\Lambda) \sim N(.1, 50^2)$, truncated at zero. Solid curves are computed under exact posterior computation, while histograms were constructed using $120,000$ posterior samples with the EF-algorithm. Prior distributions are shown by dashed curves.

Figure 6.1 shows, as the forest size increases, close agreement between exact calculations and those obtained by simulation. Under the single tree case, the EF

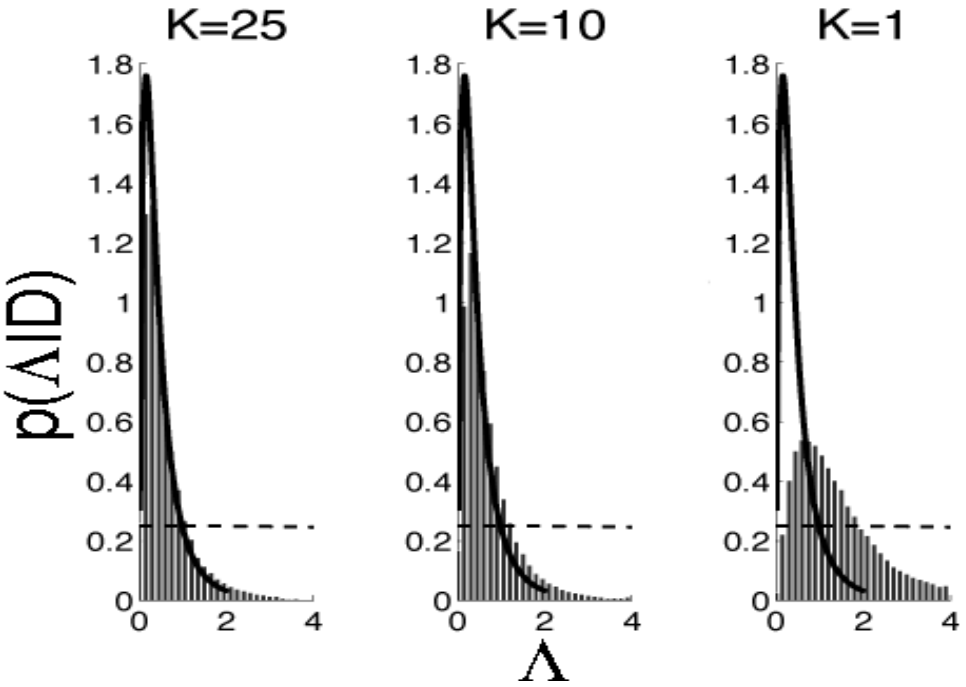

Figure 6.1: Comparison for conditional Λ posteriors on the data set for D. Persimilis and D. Bogotona (13 samples, 13 samples), with $\boldsymbol{n} = (16, 6, 8, 2, 0, 0, 0)$. $(\theta_0, \theta_1, N2)$ are evaluated at their maximum likelihood values which were found under simulation. The histogram represents draws from the posterior distribution using the EF algorithm, where the forest size has been set to $K = (1, 10, 25)$ respectively. Imposed on the simulated posteriors are the exact conditional posterior distributions. Prior distributions are denoted by dashed curves.

algorithm fails to converge after 120,000 MCMC iterations, and both the modal estimate and overall shape deviate from the true distribution. This is due to the fact that, with a single tree in the forest, tree updates occur infrequently. This irregular updating limits the exploration of the space and ultimately impedes convergence in the parameter space.

For the θ_2 parameter, we have computed the posterior distribution using the data set consisting of 13 sequences from D. Persimilis and 19 samples from D. Psuedoobscura, with $\boldsymbol{n} = (16, 5, 65, 0, 0, 1, 0)$. Figure 6.2 shows a comparison using the

EF algorithm with $K = \{1, 5, 50, 100\}$. The trend depicted, by increasing forest

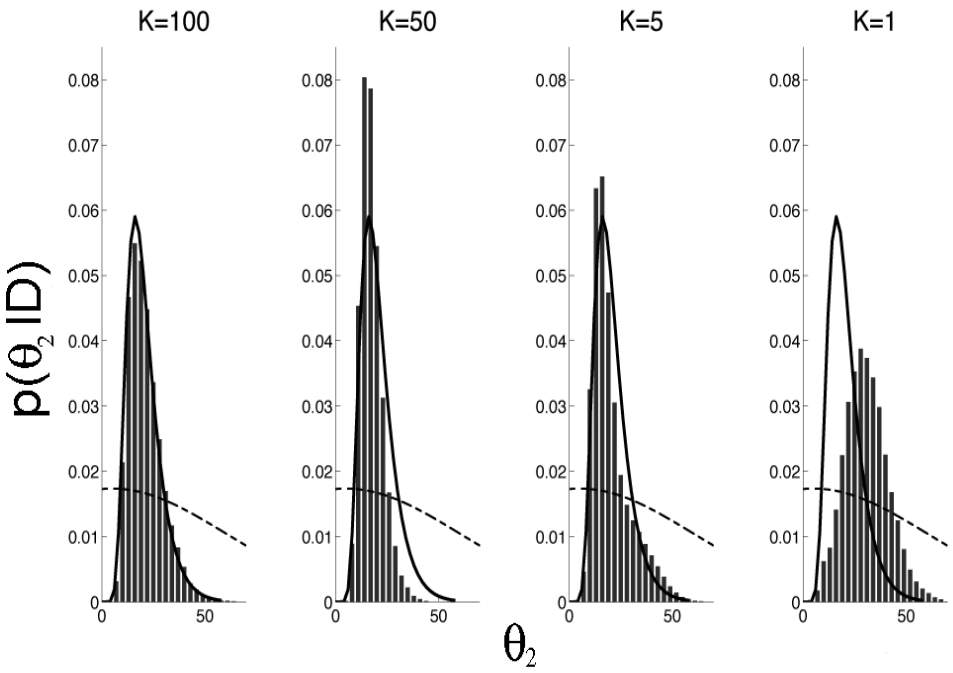

Figure 6.2: Comparison For conditional θ_2 posteriors on the data set for D. Persimilis and D. Psuedoobscura (13 samples, 19 samples), with $\boldsymbol{n} = (16, 5, 65, 0, 0, 1, 0)$. $(\theta_0, \theta_1, \Lambda)$ are evaluated at their maximum likelihood values which were found under simulation. The histogram represents draws from the posterior distribution using the EF algorithm, where the forest size has been set to $K = (1, 50, 50, 100)$ respectively. Imposed on the simulated posteriors are the exact conditional posterior distributions. Prior distributions are denoted by dashed curves.

size, shows similar patterns as the study on the smaller data set. That is, by using 1 tree, the simulation performs quite poorly. As we increase the forest size to 100 tress, we closely approximate the exact posterior distribution. Computation time for the exact curve took over 4 months, while the simulated curve using 50 trees took approximately 4 hours. Unfortunately, using the exact recursion to perform inference on the full joint posterior is not permissible since computation time for all 4-dimensions will require approximately $4^4/12 = 21.33$ years. However, sam-

pling the entire $\boldsymbol{\theta}$ vector using the EF framework requires only marginal increases in computation time as the number of estimated parameters increases, and for this case study still only requires approximately 4 hours for convergence on the full joint posterior.

We notice in a comparison of marginal posterior distributions for Λ and θ_2 in figures 6.1 and 6.2, that when the parameter reflects a high level of uncertainty, increasing the number of trees in the forest is essential for convergence to the correct target distribution. This is due to the fact that the number of trees with substantial probability weights increases as the parameters variability increases. Performance using few trees is hindered by the low acceptance rate and slows the exploration of the tree space. As a result, the parameter values proposed will only reflect those trees which are sampled. Hence, it is necessary to have the Markov chain mix efficiently in all dimensions for convergence to the correct distribution.

6.6 Computational Performance

Both figures 6.1 and 6.2, based on 120,000 sample iterations, show an increasing agreement between the exact and estimated posterior distribution, as the forest size increases. For both examples, we see that the single tree forests do not converge to the exact marginal distribution in the alloted time. This is due to the slow exploration of tree space, which contributes to the fact that small forests will require more sample iterations for convergence. For the *Dpe/Dpp* data set, using a forest size of ($K = 100$), posterior draws converged to the exact marginal distribution in only 120,000 iterations; however, for the single tree case ($K = 1$), we require about 50 million sample draws for convergence (approximately 416

times more iterations). While, the increase in the rate of convergence is problem specific, as the landscape of tree space becomes more irregular, then it is expected that the increase in forest size will be more advantageous.

For θ_2, the parameter with the widest posterior distribution, computation of the exact posterior (solid curve in figure 6.2) took over four months using the exact recursion (Uyenoyama and Takebayashi, 2004). Using forests of 50 trees, the EF results required approximately four hours (as computed using a Macintosh Dual 2.5 GHz PowerPC G5). Using the exact recursion to compute the posterior distribution for all four parameters would require on the order of $4^4/12 = 21.33$ years (as computed on a uniform grid). In contrast, using the EF algorithm, computation time remains approximately constant (4 hours) as the number the parameter space increases in dimensionality.

As the forest size increases, the only additional time penalty is in the evaluation of the posterior probability. This increases linearly as the forest size increases, since a forest size of K will require K posterior evaluations at the tree level.

6.6.1 Comparison to a previously studied importance sampling procedure

Previously, we analyzed the data set in table 5.1 using an importance sampling (IS) approach for constructing the probability of the data at a fixed parameter set $\boldsymbol{\theta}$. Although results are comparable under both the EF and IS algorithms, the EF algorithm obviates much of the preprocessing involved in the IS algorithm. The forms of the proposal distributions under both methods are the same, however

the IS method uses static driving parameters which require substantial tuning. The optimal driving set in the IS method is the maximum likelihood set but, a priori, the MLE is unknown. To address this problem, we required a two phase mode searching procedure. The first phase uses a random searching procedure for finding seeding values, for which a second phase gradient-descent algorithm is initialized. This procedure typically requires about 1-2 weeks for convergence. After this proposal tuning phase, the IS method gives the likelihood values for $\boldsymbol{\theta}$ on a grid and interpolating splines are then used in constructing a smooth likelihood surface. The results shown in figure 6.1 are comparable to those shown in figures 5.3 and 5.4. For these examples, the EF method required approximately 4 hours for convergence, while the IS method required 9 hours for convergence, in addition to a 2 week tuning period.

While this importance sampling algorithm requires many preprocessing steps, the EF method is almost fully automatic. Since the EF algorithm constantly updates the parameter values $\boldsymbol{\theta}$, the EF algorithm is not subject to the sensitivity of initial static driving values. Therefore, pre-tuning steps on the driving values are not needed. Also, the EF algorithm constructs samples from the posterior distribution $p(\boldsymbol{\theta}|D)$, so the additional burden of using interpolating splines is unnecessary.

We have found that the EF algorithm scales better to larger problems, since the memory (RAM) demands are much smaller than in our IS application. This is due to the fact that the EF algorithm only stores a small number of trees at each step of the procedure, whereas in our implementation of the IS method we must store an enormous number of trees for reliable likelihood evaluation.

6.6.2 Comparisons to IM

The IM software package (Hey and Nielsen, 2004) encompasses a closely related method to ours for parameter estimation, although one major difference is the simulation mechanism. While our method relies on the evolutionary forest for moving through the space of genealogies, IM uses parallel tempering (Geyer, 1991) to aid moves through this space. A comparison of the EF method to IM was performed on the two data sets included in Table 3.1. Parameter estimates between the two models are comparable, however our model relates the speciation event through the rate paramter Λ, where as, IM explicitly estimates the speciation time τ. We correspond their parameter $u\tau$ to our parameter $u/\lambda = \Lambda^{-1}$, since $E[\tau] = \lambda^{-1}$. Credible intervals for their τ parameter should not be compared to those of our Λ parameter, since the relationship can only be made in expectation under our modeling assumptions, which IM does not share. All the remaining parameters are immediately comparable.

We ran the IM software with 3 parallel tempering chains to aid mixing, while other input values were set to their default values. IM relies on the specification of bounded uniform prior distributions, where we defined the support on the regions as $\tau \in (0, 100]$ and $\theta_i \in (0, 300]$ $(i = 0, 1, 2)$. For the EF method we set $K = 50$ and specified the diffused prior distribution $\Lambda \sim N(.1, 30^2)$ and $\theta_i \sim N(10, 100^2)$ $(i = 0, 1, 2)$. Convergence for both methods were assessed by visual inspection of trace plots.

For the smaller data set *Dpe/Dpb* in Table 5.1, IM required 4 days for convergence, however, convergence was somewhat weak. Although the trace plots

suggest that convergence had been obtained, analysis of subsets of the posterior samples showed clear differences in the marginal distributions. Running the chain for 7 additional days did not resolve these differences, so we finally determined that convergence had been obtained to the maximum degree that the software would allow for. The EF algorithm required approximately 4 hours for convergence, where random walk proposal standard deviations were set to $(1, 3, 3, 3)$ for updating $(\Lambda, \theta_0, \theta_1, \theta_2)$, respectively. Table 6.3 shows the estimated parameter values under both the EF and IM methods. Estimates for the $\boldsymbol{\theta}$ parameters under both methods appear to agree closely. However, our estimate for Λ is slightly higher than that of IM, though the marginal deviation in these estimates are not significant.

	Λ	θ_0	θ_1	θ_2
posterior mode (EF)	0.19	2.5	3.4	4.0
95% credible intervals (EF)	(0.05,1.2)	(0.5, 51.3)	(1.7, 10.0)	(2.0, 11.0)
poserior mode (IM)	.14	2.4	2.8	5.6
95% credible intervals (IM)		(0.39, 44.7)	(1.4, 8.4)	(3.0, 21.5)

Table 6.3: Parameter estimates for the *D. persimilis* and *D. pseudoobscura bogotana* data set under the EF and IM methods. Posterior inferences based on 1,000,000 samples.

For the larger data set *Dpe/Dpp*, we ran the IM software for three weeks and were unable to obtain convergence. Two related issues appear to be behind the inability to estimate the posterior distribution within the time permitted. The first of these is that the acceptance rate on various parameters, namely τ and the genealogy itself, tended to be low. Tree updates, at times, occurred at rates of less that 1 in 10,000 (and up to 11%), which is quite low and requires that the algorithm must run for an extremely long time before convergence is obtained. Also, parameters with a large degree of uncertainty (θ_2, see figure 6.2) require a

large portion of tree space for estimation, so if tree space is explored slowly, then convergence in these parameters will suffer. Unfortunately, all of the parameters in the model are highly dependent and should not be reported if a single parameter fails to converge. Previously, Hey and Nielsen (2004) reported point estimates (scaled to our units) as ($\Lambda = 0.15$, $\theta_0 = 1.1$, $\theta_1 = 2.4$, $\theta_2 = 21.6$). A marginal assessment of these values show non statistically significant departures from our estimates (Table 6.4), which were obtained in approximately 5 hours with the EF algorithm.

	Λ	θ_0	θ_1	θ_2
Posterior Modes (EF)	0.21	5.4	4.2	15.4
95% CIs (EF)	(0.05, 1.4)	(0.9, 27.9)	(2.3, 11.1)	(9.4, 54.4)

Table 6.4: Parameter estimates and credible intervals for the *D. persimilis* and *D. pseudoobscura pseudoobscura* data set under the EF. Posterior inferences based on 1,000,000 samples.

6.7 Acceptance probability

To investigate further the implications of forest size, we examined the acceptance probability of moves in tree space. For the *D. persimilis* and *D. pseudoobscura pseudoobscura* data set (*Dpe/Dpp* column in Table 5.1), we fixed the population parameter values at their maximum likelihood values ($\Lambda = 0.12$, $\theta_0 = 0.81$, $\theta_1 = 2.71$, $\theta_2 = 18.21$) so that MCMC moves are limited to the forest space. Figure 6.3 shows an increase in the acceptance rate with forest size, confirming that larger forests promote more rapid exploration of the tree space and convergence of the full joint posterior distribution over both trees and parameters. Expansion of forest size reduces the effect of the substitution of a single tree, permitting moves

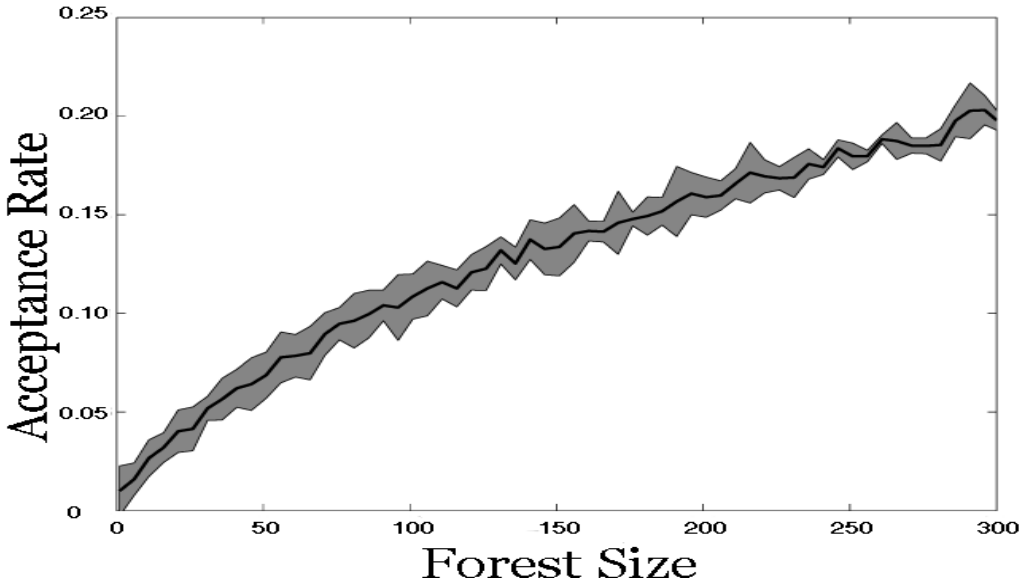

Figure 6.3: Acceptance rate increases with forest size. Black line: mean acceptance rate from six MCMC runs, each comprising 10,000 iterations for each point. Grey region: 95% pointwise confidence interval around the mean for the six runs.

to occur more frequently. For arbitrarily large forests,

$$\lim_{K\to\infty} \frac{q(D, f^{(c)}|\boldsymbol{\theta}^{(i)})}{q(D, f^{(i-1)}|\boldsymbol{\theta}^{(i)})} = 1, \tag{6.11}$$

indicating that in the ratio of the proposal distributions

$$\frac{q(f^{(c)}|D, \boldsymbol{\theta}^{(i)})M(t^c, f^{(c)})g(t_j|\boldsymbol{\theta}^{(i)}, D)}{q(f^{(i-1)}|D, \boldsymbol{\theta}^{(i)})M(t_j, f^{(i-1)})g(t^c|\boldsymbol{\theta}^{(i)}, D)},$$

$M(t, f)g(t|\boldsymbol{\theta}, D)$ will dominate the Metropolis-Hastings ratio (6.8). The rate of updates in the space of histories is usually a limiting factor in evolutionary inference, so being able to increase this rate by simply changing the size of the forest space greatly increases the overall convergence rate for all the parameters in the model. Therefore, increasing the acceptance rate in forest space will ultimately

have the effect of increasing the convergence rate in the population parameter space.

6.8 Conclusion And Discussion

In this study, we have developed a novel approach to efficiently sample from the joint posterior distribution of population parameters in the coalescence framework. While, a natural sampling technique is to sample from the space where the coalescence tree is one of the marginal parameters of interest, these methods must make use of cumbersome proposal distributions, where convergence can often times be a limiting factor. The Evolutionary Forest algorithm makes use of augmenting the population parameters with a forest of possible genealogical histories. Through this construction, proposal distributions on the individual trees may be constructed in an independent manner, while maintaining efficient mixing in both parameter space and forest space. The EF method constructs credible intervals within several minutes, obviating the need for computationally demanding methods in which profile likelihood intervals are considered.

6.8.1 Decoupling The Parameter Space From The Genealogy

In appendix A we placed some importance on being able to separate the parameter space $\boldsymbol{\theta}$ from the genealogy it self. That is, all demographic parameters of interest were not to be explicitly rooted within the tree, but rather were to be expressed in terms of rates that drive the stochastic process. It is due to this modeling

feature that we are able to apply the EF algorithm and still perform inference in all the parameters of interest. If, for example, the speciation event were rooted within the tree, then under the EF framework, only the margins for $\{\theta_0, \theta_1, \theta_2\}$ would converge to the correct distribution. However, the divergence time will not converge since it is inherently coupled with the tree. Since the instrumental trees within the marginal forests do not converge to the same distribution as under the original $p(\boldsymbol{\theta}, t|D)$ distribution, inference cannot be performed on the trees resulting from EF. However, by separating out all of the parameters of interest, we can apply EF, which converges rapidly and still retains all the inferential ability as specified under the single tree distribution.

Chapter 7

Conclusions and Future Work

In this thesis we have introduced a new sampling tool called the Evolutionary Forest (EF) algorithm. This algorithm is a powerful device for searching around complicated spaces and ultimately yields the population parameters that drive the coalescence process. While, we have chosen in this thesis to only focus on evolutionary models, where the stochastic realizations are bifurcating trees, we can apply this algorithm to any arbitrary process and retain the driving parameters of the process.

Since the EF algorithm has shown dramatic speed gains over traditional methods, more complex models can be developed.

Coalescence models in which more that two species are of interest will be developed. This development will provide a bridge between the phylogenetic and population biology communities, since we will be able to reconstruct species level phylogenetic trees from gene level stochastic processes. To date, we are unaware

of any groups successfully achieving this goal.

Multi locus analyses are often desired since multiple regions of the genome can provide more information about the population parameters than a single locus. While, these parameters are generally common to each locus in the genome, the genealogy is not the same. Hence a full multi locus analysis demands that multiple dependent genealogies are modeled simultaneously. Even though the EF algorithm greatly reduces the computational burden induced by the coalescence framework, multi locus analyses over a *hand full* of loci will be troublesome. For this reason, we envision deploying the EF algorithm with a resampling approach to building multiple joint posterior distributions from only subsets of the data. Averaging over large quantities of sub-sampled posterior distributions will eventually yield the overall posterior distribution. The key to this procedure is that each sub-posterior must be generated quickly, which the EF algorithm already accomplishes.

In the future, we will extend the EF algorithm to more arbitrary processes and develop it as a general sampling tool. In particular, we will show how this methodology can be used for sampling from multidimensional multi modal distributions in which only the simple margins are of interest. After which, we will extend the theory to cases in which all margins are of interest.

In model selection problems, the reversible jump algorithm has often been employ with moderate success. It has been our experience, that in large model spaces, proposing model jumps so that between high probability models is often problematic. Again proposal distributions are at the heart of the sampling difficulty in

model selection problems. In future applications, we will extend the EF algorithm to problems in which finding *the best* model (or set of models) is of interest. In the model selection context, the EF algorithm will rely on making moves between forests of models in conjunction with intermediate reversible jump moves.

Future applications will also include time series models for modeling the spread of infectious disease. These models will show by inclusion of genetic (haplotype) data from the pathogen, detection of the infection rate will be massively amplified. This approach will use hierarchical branching processes for modeling the birth death process of the disease. For controlling for seasonal effects and the frequency of the haplotypes in the population fourier harmonics and dirichlet mixtures will be invoved.

Appendix A

Recursion In Probability Generating Functions

A.1 Recursion In Probability Generation Functions

We begin by describing the basic structure of all of the speciation models that will be studied in this thesis. We will derive the likelihood of the observed mutational data given in table 3.1 from the probability generating function (pgf) of the model. Letting $g_l(\boldsymbol{a})$ be the pgf for the coalescence process at level l of the genealogy, where $\boldsymbol{a} = (a_1, a_2, a_3, a_4, a_5, a_6, a_7)$ are the pgf parameters corresponding to the seven types of segregating sites (3.1), we can write a recursion in pgfs from level $l-1$ to level l as

$$\boldsymbol{g}_l(\boldsymbol{a}) = \boldsymbol{F}_l(\boldsymbol{a})(\boldsymbol{U}_l\boldsymbol{g}_l(a) + \boldsymbol{V}_l\boldsymbol{g}_{l-1}(\boldsymbol{a})). \tag{A.1}$$

$\boldsymbol{U}_l$ is a square matrix which reflects within level transitions, which change the structure on the tree but preserve the level of the tree. These within level transitions can reflect any biological process of interest which doesn't reduce the tree level, for example: speciation, migration, and recombination are within level tran-

sitions. $\boldsymbol{V}_l$ is a rectangular matrix which reflects coalescence, taking the level from l to $l-1$. This gives the fundamental structure for all the coalescence models which we explore in this thesis. That is a within level transition or coalescence transition may occur, with mutational events occurring in between. $\boldsymbol{F}_l(\boldsymbol{a})$ is a square matrix consisting of pgfs for the mutational process of mutation counts $\boldsymbol{a}$ on level l.

Since either a within level transition of a coalescence event occurs with mutational events in between transitions, we have

$$\boldsymbol{U}_{l,\alpha} + \boldsymbol{V}_{l,\alpha} = 1,$$

where $\boldsymbol{U}_{l,\alpha}$ and $\boldsymbol{V}_{l,\alpha}$ denote the α row sums of the matrices $\boldsymbol{U}_l$ and$\boldsymbol{V}_l$ respectively.

The recursion in pgfs given by equation (A.1) factors into the components $\boldsymbol{F}_l(\boldsymbol{a})$ and $(\boldsymbol{U}_l\boldsymbol{g}_l(a) + \boldsymbol{V}_l\boldsymbol{g}_{l-1}(\boldsymbol{a}))$ since given the level l, the mutational process and the transition process are independent processes, and pgfs of independent processes factor into their respective pgfs.

Since mutations from the most recent common ancestor don't segregate in the sample, we have the initial condition as

$$\boldsymbol{g}_1(\boldsymbol{0}) = \boldsymbol{1}.$$

Rearranging terms, we can rewrite recursion (A.1) as

$$\boldsymbol{g}_l(\boldsymbol{a}) = [\boldsymbol{I} - \boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{U}_l]^{-1}\boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{V}_l\boldsymbol{g}_{l-1}(\boldsymbol{a}). \tag{A.2}$$

which more simply describes how transitions are made from level $l-1$ to l.

On each level of the tree a maximum of three mutational types may coexist. We write the counts of these mutations as (c_1, c_2, c_3), where

$$c_1 = \begin{cases} a_1 & \text{if } l_1 > 1 \text{ or } l_3 > 0 \\ a_2 & \text{if } l_1 = 1 \text{ and } l_3 = 0 \end{cases}$$

$$c_2 = \begin{cases} a_3 & \text{if } l_2 > 1 \text{ or } l_3 > 0 \\ a_4 & \text{if } l_2 = 1 \text{ and } l_3 = 0 \end{cases} \tag{A.3}$$

$$c_3 = \begin{cases} a_5 & \text{if } l_1, l_2 > 0 \text{ or } l_3 > 1 \\ a_6 & \text{if } l_1 = 0,\ l_2 > 0, \text{ and } l_3 = 1 \\ a_7 & \text{if } l_1 > 0,\ l_2 = 0, \text{ and } l_3 = 1. \end{cases}$$

Hence, if the mutation type a_1 (segregation/absent) is placed at level l, it is not possible to place an a_2 (fixed/absent) mutation type on the same level. From this, it maybe useful to only discuss to map the seven mutation types to the 3-vector of mutations (c_1, c_2, c_3) on level l Note however that the state space on level l, $\boldsymbol{S}_l$ must be know to distinguish the types of the mutations (c_1, c_2, c_3).

In general, a coalescence model can be viewed as a structured Poisson process (cite something, maybe chapter 1). The waiting times until jth event type, where the event types are: mutations, within level transitions, or coalescence events, are competing exponential distributions. The probability that the first event to happen is of type j is $\gamma_j / \sum_i \gamma_i$, where γ_j is the exponential parameter for the corresponding event. Hence the only components that actually appear in the model are the rates the components in the Poisson process. We can re-parameterize equation (A.2) in terms of the rate parameters from the competing exponential

terms by considering the rate matrices

$$\boldsymbol{P}_l \qquad \text{and} \qquad \boldsymbol{Q}_l,$$

where $\boldsymbol{P}_l$ is a square matrix with elements $p_{i,j}$ being the within level transition rate of moving from the ith state to the jth state. $\boldsymbol{Q}_l$ is a rectangular matrix with elements $q_{i,j}$ representing the rate of coalescing from state i to state j. Letting $\boldsymbol{C}_l$ be a diagonal matrix with diagonal entries

$$\boldsymbol{C}_{l,\alpha} = \boldsymbol{P}_{l,\alpha} + \boldsymbol{Q}_{l,\alpha},$$

we have

$$\begin{aligned} \boldsymbol{U}_l &= \boldsymbol{C}_l^{-1}\boldsymbol{P}_l \\ \boldsymbol{V}_l &= \boldsymbol{C}_l^{-1}\boldsymbol{Q}_l. \end{aligned}$$

A.1.1 Mutational Process

The probability that a transition event occurs before a mutation event follows as

$$p_{l,\alpha} = \frac{\boldsymbol{C}_{l,\alpha}}{\boldsymbol{C}_{l,\alpha} + \mu l},$$

since there are l branches on level l and a mutation can fall with equal probability on each branch. It follows that the number of mutations ($N_{l,\alpha}$) accumulated in state α follows a geometric distribution with parameter $p_{l,\alpha}$, so that

$$P(N_{l,\alpha} = k) = (1 - p_{l,\alpha})^k p_{l,\alpha}.$$

These k mutational types can occur on any of the 3 available branch type $i \in (1, 2, 3)$ with probabilities

$$\begin{aligned} p_1 &= l_1/l \\ p_2 &= l_2/l \\ p_3 &= l_3/l \end{aligned}$$

where l_i denotes the number of branches on level l of type i. Therefore, given $N_{l,\alpha} = k$, the number of segregating sites accumulated in state α, the distribution of number of mutations that arose on type 1,2, and 3 branches (x,y,and z respectively) occurs with probability

$$\frac{k!}{x!y!z!} p_1^x p_2^y p_3^z. \tag{A.4}$$

From this, we can derive the pgf for the configuration of mutations $\boldsymbol{a} = (a_1, a_2, a_3, a_4, a_5, a_6, a_7)$ occurring on level l corresponding to the elements $f_{l,\alpha}(\boldsymbol{a})$ of the matrix $\boldsymbol{F}_l(\boldsymbol{a})$, where $\boldsymbol{F}_l(\boldsymbol{a})$ is the diagonal matrix mutation matrix and $f_{l,\alpha}(\boldsymbol{a})$ are its diagonal entries.

Since the joint distribution for the total number of mutational events $N_{l,\alpha} = k$ is distributed geometrically, and the arrangement of the k mutations on type 1,2, and 3 branches has a multinomial distribution, the joint probability distribution follows as

$$\left(\frac{k!}{x!y!z!} p_1^x p_2^y p_3^z \right) \left((1 - p_{l,\alpha})^k p_{l,\alpha} \right). \tag{A.5}$$

Since only 3 possible mutation types can arise on each level, corresponding to type 1,2, and 3 branches, the joint probability generating function, for the sampling

distribution in (A.5), follows as

$$f_{l,\alpha}(\boldsymbol{a}) = \frac{\boldsymbol{C}_{l,\alpha}}{\boldsymbol{C}_{l,\alpha} + u[l_1(1-c_1) + l_2(1-c_2) + l_3(1-c_3)]},$$

where the type assignment of (c_1, c_2, c_3) depends on the branch types and can be determined from (A.3). Therefore, depending on the state configuration $\boldsymbol{S}_l$, the full matrix of pgfs for the mutational process $\boldsymbol{F}_l(\boldsymbol{a})$ can be constructed with $f_{l,\alpha}(\boldsymbol{a})$ as the diagonal entries. All the others entries remain zero since the mutational process doesn't change the state configuration.

A.1.2 Complete Tree PGF

It is often convenient to think about the recursion in pgfs through the rate matrices $\boldsymbol{P}_l$ and $\boldsymbol{Q}_l$ instead of the transition matrices $\boldsymbol{U}_l$ and $\boldsymbol{V}_l$, so rewriting the recursion in pgfs given by equation (A.2) with the corresponding rate matrices gives the recursion

$$\boldsymbol{g}_l(\boldsymbol{a}) = [\boldsymbol{I} - \boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{P}_l]^{-1}\boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{Q}_l\boldsymbol{g}_{l-1}(\boldsymbol{a}), \tag{A.6}$$

where

$$\boldsymbol{D}_l(\boldsymbol{a}) = \boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{C}_l^{-1}.$$

Given the initial boundary condition $\boldsymbol{g}_l(\boldsymbol{0}) = \boldsymbol{1}$, the total probability generating function for the entire tree results from the product across all levels of the tree

$$\boldsymbol{g}_L(\boldsymbol{a}) = \prod_{l=2}^{L}[\boldsymbol{I} - \boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{P}_l]^{-1}\boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{Q}_l, \tag{A.7}$$

where L is the total number of samples in the data.

We note that the size of the matrices in the product given by equation (A.7) vary

in size, depending on the level l. This is since the size of the state space in the middle of the tree have the largest number of possible attainable states, which is determined by the initial sample. On levels near the boundaries of the tree (top or bottom), the number of attainable states decrease, and the size of the transition matrices decreases (see figure A.1). The line moving through the state space

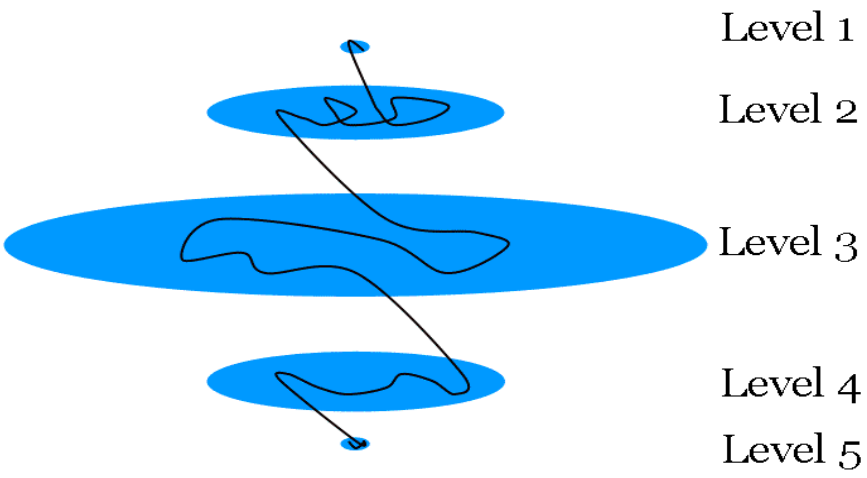

Figure A.1: The circles represent the size of the attainable state space on the corresponding level. The line moving through the levels represents the within level and between level transitions

.

represents a possible path taken by transitioning within levels and between levels. From equation (A.2), we can see by ignoring the mutational process ($\boldsymbol{F}_l(\boldsymbol{a}) = \boldsymbol{I}$), the transition probability is governed by the term

$$(\boldsymbol{I} - \boldsymbol{U}_l)^{-1}\boldsymbol{V}_l = (\sum_{i=0}^{\infty} \boldsymbol{U}_l^i)\boldsymbol{V}_l \tag{A.8}$$

which reflects that as many within level transitions may occur between coalescence transitions, so long as the powers of $\boldsymbol{U}_l$ remain positive.

A.1.3 Computing The Exact Likelihood

Given the observed mutations in the sample $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5, n_6, n_7)$, in order to compute the likelihood function of the data as a function of the rate parameters controlling the transition processes, we need to take the 7-fold derivative of the product of pgfs given in equation (A.7). This creates a recursion in the 7-fold derivatives as well as level numbers, which is massively computationally intensive for even relatively small data sets. We derive the recursive form of the exact likelihood function in only the simplest classes of models, where the matrix $\boldsymbol{P}_l$ is nilpotent. That is, we have

$$\boldsymbol{P}_l^k = 0 \qquad \text{for } k \geq 2.$$

This simplification directly addresses the computational burden of dealing with that matrix inversions in equation (A.7), since

$$[\boldsymbol{I} - \boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{P}_l]^{-1} = \boldsymbol{I} + \boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{P}_l. \tag{A.9}$$

In this specific case, we allow for only one possible within level transition between coalescence events. While this may seem like an artificial constraint, there is one very important class of models where this assumption is justified. That is the case when the only within level transition event is a speciation event. Since a species split can only occur once, for the two species problem, in the entire history of genes, $\boldsymbol{P}_l$ is in fact nilpotent since only one within level transition is permitted between coalescence events.

Now writing equation (A.6) as

$$\boldsymbol{g}_l(\boldsymbol{a}) = \boldsymbol{R}_l(\boldsymbol{a})\boldsymbol{g}_{l-1}(\boldsymbol{a}), \tag{A.10}$$

where

$$\boldsymbol{R}_l(\boldsymbol{a}) = [\boldsymbol{I} + \boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{P}_l]\boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{Q}_l,$$

we find the probability of observing the mutation array $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7)$ by recursively applying the product rule for derivatives to equation (A.10) and obtain

$$\frac{\boldsymbol{g}_l^{(\boldsymbol{p})}(\boldsymbol{0})}{\prod_{i=1}^{7} p_i!} = \sum_{\boldsymbol{q}} \frac{\boldsymbol{R}_l^{(\boldsymbol{q})}(\boldsymbol{0})}{\prod_{i=1}^{7} q_i!} \frac{\boldsymbol{g}_{l-1}^{(\boldsymbol{p}-\boldsymbol{q})}(\boldsymbol{0})}{\prod_{i=1}^{7}(p_i - q_i)!}. \tag{A.11}$$

This expression in derivatives represents the $\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5, q_6, q_7)$ mutations that occur on level l and the $\boldsymbol{p} - \boldsymbol{q}$ mutations that occurred previous to level l (see figure A.2), where the sum over $\boldsymbol{q}$ runs over all possible arrangements of mutations on level l, and $\boldsymbol{p} - \boldsymbol{q}$ mutations on older branches in the history (see figure A.2).

A.1.4 Recursion Derivatives

Differentiating $\boldsymbol{R}_l(\boldsymbol{a}) = [\boldsymbol{I} + \boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{P}_l]\boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{Q}_l$ requires differentiation of the $\boldsymbol{D}_l(\boldsymbol{a})$ matrix only, since this is the only matrix dependent on $\boldsymbol{a}$. An analysis of the element by element differentiation of this matrix yields

$$\frac{d\boldsymbol{D}_l(\boldsymbol{a})}{da_i} = \boldsymbol{D}_l(\boldsymbol{a})^2 \boldsymbol{E}_{l,i},$$

where $\boldsymbol{E}_{l,i}$ is a diagonal matrix with diagonal elements being one of $(\mu l_1, \mu l_2, \mu l_3)$. Determination of the diagonal elements in $\boldsymbol{E}_{l,i}$ depends on the level configuration given by (A.3) and the index i. For example if the index $i \in 1, 2$ and the state

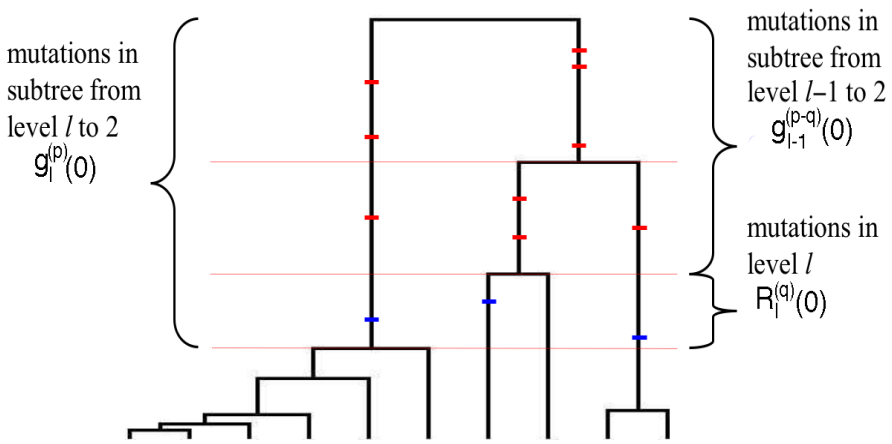

Figure A.2: Equation (A.11) represents the $\boldsymbol{q}$ mutations that occur on level l and the $\boldsymbol{p-q}$ mutations that occurred previous to level l

.

satisfies either of the first two conditions given by (A.3), the diagonal entry for that state is μl_1, otherwise it is 0. For $i \in 3, 4$, the diagonal elements of $\boldsymbol{E}_{l,i}$ will either be μl_2 or 0, and similarly for $i \in 5, 6, 7$, the diagonal elements of $\boldsymbol{E}_{l,i}$ will either be μl_3 or 0. We have that

$$\begin{aligned}(\boldsymbol{D}_l(\boldsymbol{a})\boldsymbol{E}_{l,i})^k &= \boldsymbol{D}_l(\boldsymbol{a})^k \boldsymbol{E}_{l,i}^k \\ &= \boldsymbol{E}_{l,i}^k \boldsymbol{D}_l(\boldsymbol{a})^k,\end{aligned}$$

since $\boldsymbol{D}_l(\boldsymbol{a})$ and $\boldsymbol{E}_{l,i}$ are diagonal matrices. Also, the following relationship holds

$$\boldsymbol{P}_l \boldsymbol{E}_{l,i} = \boldsymbol{E}_{l,i} \boldsymbol{P}_l.$$

The first derivative of $\boldsymbol{R}_l(\mathbf{0})$ follows as

$$\begin{aligned}\frac{\partial \boldsymbol{R}_l(\mathbf{0})}{\partial a_i} &= [\boldsymbol{I} + \boldsymbol{D}_l(\mathbf{0})\boldsymbol{P}_l]\boldsymbol{D}_l(\mathbf{0})^2\boldsymbol{E}_{l,i}\boldsymbol{Q}_l + \boldsymbol{D}_l(\mathbf{0})^2\boldsymbol{E}_{l,i}\boldsymbol{P}_l\boldsymbol{D}_l(\mathbf{0})\boldsymbol{Q}_l \\ &= \boldsymbol{D}_l(\mathbf{0})\boldsymbol{E}_{l,i}[\boldsymbol{I} + \boldsymbol{D}_l(\mathbf{0})\boldsymbol{P}_l + \boldsymbol{P}_l\boldsymbol{D}_l(\mathbf{0})]\boldsymbol{D}_l(\mathbf{0})\boldsymbol{Q}_l.\end{aligned}$$

From this calculation, the second derivative follows as

$$\begin{aligned}\frac{\partial^2 \boldsymbol{R}_l(\mathbf{0})}{\partial a_i^2} &= \boldsymbol{D}_l(\mathbf{0})^2\boldsymbol{E}_{l,i}^2[\boldsymbol{I} + \boldsymbol{D}_l(\mathbf{0})\boldsymbol{P}_l + \boldsymbol{P}_l\boldsymbol{D}_l(\mathbf{0})]\boldsymbol{D}_l(\mathbf{0})\boldsymbol{Q}_l \\ &+ \boldsymbol{D}_l(\mathbf{0})\boldsymbol{E}_{l,i}[\boldsymbol{E}_{l,i}\boldsymbol{D}_l(\mathbf{0})^2\boldsymbol{P}_l + \boldsymbol{E}_{l,i}\boldsymbol{P}_l\boldsymbol{D}_l(\mathbf{0})^2]\boldsymbol{D}_l(\mathbf{0})\boldsymbol{Q}_l \\ &+ \boldsymbol{D}_l(\mathbf{0})\boldsymbol{E}_{l,i}[\boldsymbol{I} + \boldsymbol{D}_l(\mathbf{0})\boldsymbol{P}_l + \boldsymbol{P}_l\boldsymbol{D}_l(\mathbf{0})]\boldsymbol{E}_{l,i}\boldsymbol{D}_l(\mathbf{0})^2\boldsymbol{Q}_l \\ &= 2\boldsymbol{E}_{l,i}^2[\boldsymbol{D}_l(\mathbf{0})^3 + \boldsymbol{D}_l(\mathbf{0})^2\boldsymbol{P}_l\boldsymbol{D}_l(\mathbf{0}) + \boldsymbol{D}_l(\mathbf{0})\boldsymbol{P}_l\boldsymbol{D}_l(\mathbf{0})^2 + \boldsymbol{D}_l(\mathbf{0})^3\boldsymbol{P}_l]\boldsymbol{D}_l(\mathbf{0})\boldsymbol{Q}_l\end{aligned}$$

By induction, we can show for $q = \sum_{i=1}^{7} q_i$ being the total number of accumulated mutations on level l, the higher order derivatives follow as

$$\boldsymbol{R}_l^{(\boldsymbol{q})}(0) = q!\left(\prod_{i=1}^{7}\boldsymbol{E}_{l,i}^{q_i}\right)\left(\boldsymbol{D}_l(\mathbf{0})^q + \sum_{j=0}^{q}\boldsymbol{D}_l(\mathbf{0})^{j+1}\boldsymbol{P}_l\boldsymbol{D}_l(\mathbf{0})^{q-j}\right)\boldsymbol{D}_l(\mathbf{0})\boldsymbol{Q}_l.$$

Therefore, in the case where $\boldsymbol{P}_l$ is nilpotent, we have a relatively simple expression to evaluate at each step in recursion (A.11).

By iterating over all over all mutation configurations on level l, we can built up the probability of observing every configuration of mutations at the next level in the tree $l+1$. Hence by iterating equation (A.2) over mutations followed by iterating over level number and repeating, we will eventually reach the probability of the array of mutations observed in our sample $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5, n_5, n_6, n_7)$ on level L, where L is the number of of samples. The state configuration of the

observed data is $(L_1, 0, 0|0, L_2, 0|0, 0, 0)$, for $L_1 + L_2 = L$, so the entry of $\frac{\boldsymbol{g}_L^{(\boldsymbol{n})}(\mathbf{0})}{\prod_{i=1}^{r} n_i!}$ corresponding to this state gives us the likelihood of our sample.

A.2 Conveniences And Limitations

The coalescence models studied within this thesis all have the structural form in pgfs

$$\boldsymbol{g}_l(\boldsymbol{a}) = [\boldsymbol{I} - \boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{U}_l]^{-1}\boldsymbol{F}_l(\boldsymbol{a})\boldsymbol{V}_l\boldsymbol{g}_{l-1}(\boldsymbol{a}).$$

We immediately see that the state space $\boldsymbol{S}_l$ does not depend on the mutation process. This provides us with the relatively compact state space representation

$$\boldsymbol{S}_l = (l_{1,1}, l_{1,2}, l_{1,3}|l_{2,1}, l_{2,2}, l_{2,3}|l_{0,1}, l_{0,2}, l_{0,3}),$$

on level l of the genealogy. Since the matrix size is a limiting component of computability, taking the mutational events out of the state space greatly increases computational feasibility. However, even with the states representing only the tree topology, the number of possible tree topologies is massive. As figure A.1 shows, the number of attainable states from the observed sample grows through the middle of the tree. Hence computation on the midlevels of the genealogy can take arbitrarily large amounts of time, depending on the number of samples.

The model formulation given is also independent of branch length. The likelihood is only dependent on the relative orderings of events. While the branch length can easily be imputed for a complete resolution of the tree structure, the number of mutational events between transition events is the only real information that can be provided about the event times in the tree. Hence, for inferential purposes, models without branch lengths are sufficient for parameter estimation and reduce computational burden.

The parameters within the model, controlling within level and between level transitions, are the driving components of the evolutionary process. None of the parameters of interest are actually rooted in the tree. It is conceivable to imagine the speciation event as a node fixed in the coalescent, however we model this purely as a rate parameter controlling the within level process. If any parameters were modeled as nodes within the tree, it would not be the case that branch lengths case could be ignored, since the age at these nodes would be needed. The gene tree in this context would no longer serve as a purely latent process since we would need to examine it in order to obtain the values of the rooted parameters. We will leverage the fact that under our model, the gene tree is purely a latent and discardable process and apply a particular data augmentation to the tree space which preserves the marginal distributions of the desired parameters (see chapter 6).

Bibliography

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). Integrating likelihood methods for eliminating nuisance parameters. *Stat. Sci.* **14**, 1–28.

Dobzhansky, T. and Powell, J. R. (1975). *Drosophila pseudoobscura* and its American relatives, *Drosophila persimilis* and *Drosophila miranda*. In R. C. King, ed., *Invertebrates of Genetic Interest*, 537–587. Plenum Press, New York.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* **3**, 87–112.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates, Massachusetts.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford Univ. Press, Oxford, 1st edn.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Stat. Assoc.* **85**, 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In E. M. Keramidas, ed., *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156–163. Interface Foundation of North America, Fairfax Station, VA.

Griffiths, R. C. and Lessard, S. (2005). Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theor. Pop. Biol.* **68**, 167–177.

Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–760.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics, eds., *Oxford Surveys in Evolutionary Biology*, vol. 7,

1–44. Oxford Univ. Press, New York.

Karlin, S. and McGregor, J. (1972). Addendum to the paper of W. Ewens. *Theor. Pop. Biol.* **3**, 113–116.

Kingman, J. F. C. (2000). Origins of the coalescent: 1974–1982. *Genetics* **156**, 1461–1463.

Kumar, S. and Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. (USA)* **99**, 803–808.

Leman, S. C., Chen, Y., Stajich, J. E., Noor, M. A. F., and Uyenoyama, M. K. (2005). Likelihoods from summary statistics: Recent divergence between species. *Genetics* **171**, 1419–1436.

Machado, C. A., Kliman, R. M., Markert, J. A., and Hey, J. (2002). Inferring the history of speciation from multilocus DNA sequence data: The case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol.* **19**, 472–488.

Noor, M. A. F., Grams, K. L., Bertucci, L. A., and Reiland, J. (2001). Chromosomal inversions and reproductive isolation of species. *Proc. Natl. Acad. Sci. (USA)* **98**, 12084–12088.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

Ross, S. M. (2005). *A first course in probability.* Prentice Hall, New York.

Schaeffer, S. W. (1995). Population genetics in *Drosophila pseudoobscura*: A synthesis based on nucleotide sequence data for the *Adh* gene. In L. Levine, ed., *Genetics of natural populations: The continuing importance of Theodosius Dobzhansky*, 329–352. Columbia U. Press, New York.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genet. Res.* **58**, 167–175.

Tamura, K., Subramanian, S., and Kumar, S. (2004). Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21**, 36–44.

Uyenoyama, M. K. and Takebayashi, N. (2004). A simple method for computing exact probabilities of mutation numbers. *Theor. Pop. Biol.* **65**, 271–284.

Wakeley, J. and Hey, J. (1997). Estimating ancestral population parameters. *Genetics* **145**, 847–855.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* **7**, 256–276.

Wiuf, C. and Donnelly, P. (1999). Conditional genealogies and the age of a neutral mutant. *Theor. Pop. Biol.* **56**, 183–201.

Biography

Scotland Charles Leman was born in Santa Clara, California on December 5, 1974. He graduated from The University of California, Davis with a Bachelor of Science in Mathematics in May of 2001. Following this he received a Master of Science in Scientific Computing and Computational Mathematics from Stanford University in May of 2003. It was during this time that his interest in Bayesian Statistics was developed. Scotland decided to pursue his interest in Bayesian methodology and computational science and enter the Ph.D. program at Duke university.